International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-2, March 2022
https://doi.org/10.55524/ijircst.2022.10.2.113
Article ID IRPP1053, Pages 599-605
www.ijircst.org

Customer Churn Scrutiny and Prediction Using Data Extraction Models in Funding Sectors

P. Ramalingamma[1], G. Subba Rao[2], K. Hari[3], and T. R. Chaitanya[4]

[1,2,3]Assistant Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India
[4]Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India

Correspondence should be addressed to P. Ramalingamma; ithod@pace.ac.in

ABSTRACT- A new method for customer churn analysis and prediction has been proposed. The method uses data Extraction model in Funding 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 catalog technique from data Extraction that produces a machine learning model. This study tested 5 different catalog 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— Customer churn, prediction, data Extraction, catalog, machine learning.

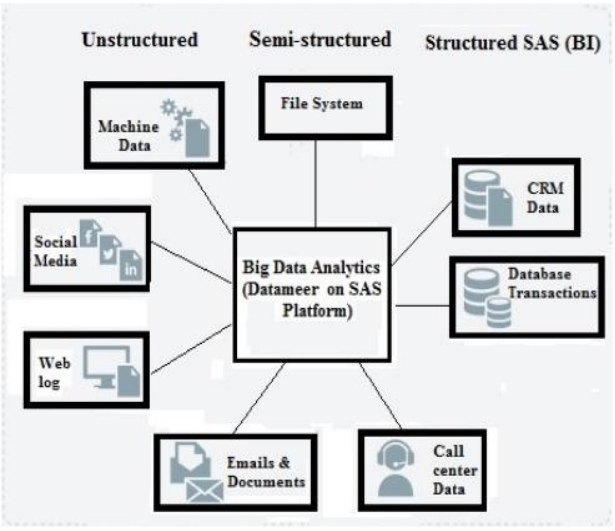

Figure 1: Flow chart of Big data Analysts

I. INTRODUCTION

We can depict the information through an image.

II. THEORETICAL FRAM

A. Customer Churn

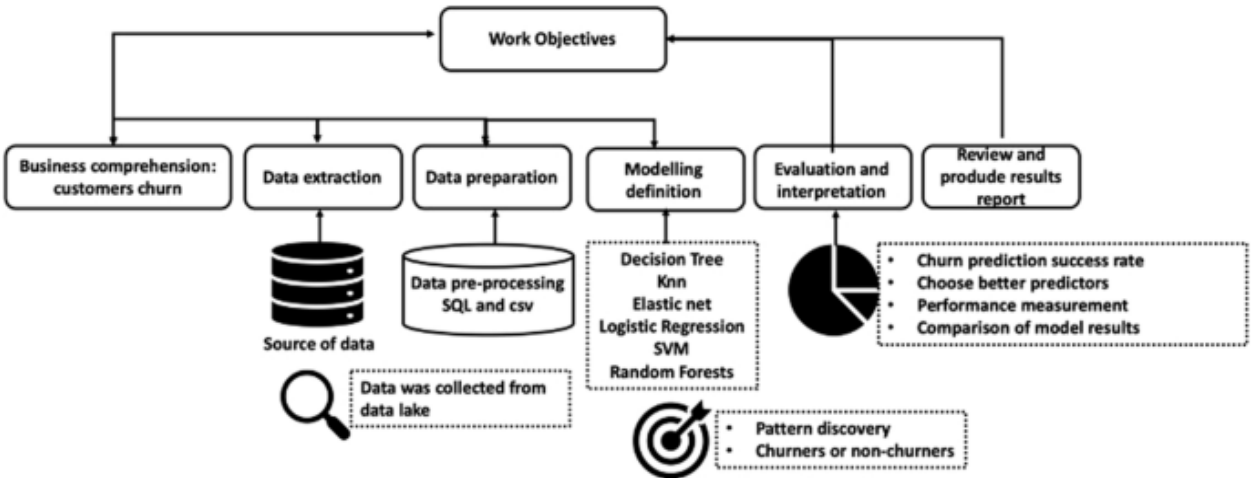

Figure 2: Bigdata Block Diagram

From the above diagram, Data Extraction will done. Evaluation of modeling can use k-fold cross validation which is one of the popular model validation methods used. This validation method works by dividing a number

of data k and repeating iterations as much as k as well. This is so that the resulting model is not only good when using training datasets but also good for other datasets (overfitting) [5],[11].

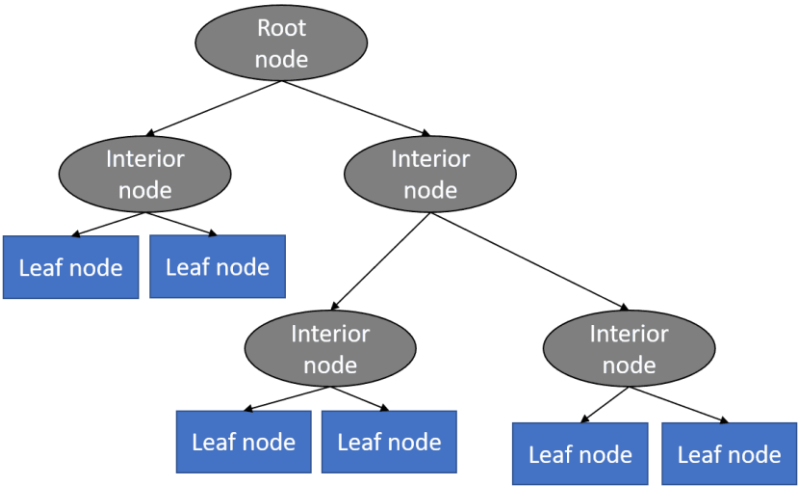

Figure 3: Flow chart Between K fold

B. Crisp-Dm

This study used C RISP-DM in conducting this research. CRISP-DM was compiled jointly by Daimler Chrysler, SPSS and NCR in 1996 and was first published in 1999 and reported as a leading methodology for data Extraction projects and analytic predictions in polls conducted in 2002, 2004 and 2007 [13],[14]. There are six phases in CRISP-DM which can be seen in Figure 4 below:

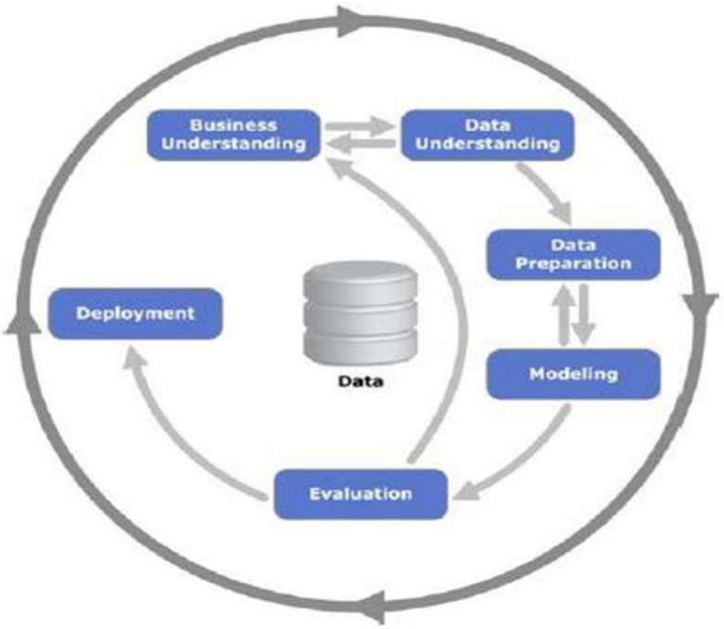

Figure 4: CRISP-DM Schema

The six phases can be explained as follows:

- Business understanding focuses on understanding the objectives of the research and the views of the business.
- Data understanding to understand the data to be used.
- Data preparation includes all activities to build the final dataset of raw data.
- Modeling for processing data using selected data Extraction methods and conducting experiments with parameters calibrated to optimal values.
- Evaluation is evaluating all the steps implemented to build the model reviewed and ensuring its business objectives are achieved.
- 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.

C. Related Work

Previous research said that the use of data Extraction can help predict customer churn. These studies use various methods in learning.

Oyeniyi & Adeyemo in 2015 examined the use of customer demographic data and customer transaction data to detect churners using the k- Means method and JRIP

algorithm. This study only uses 500 data with only four attributes. This is because this study only focuses on customers who are still carrying out transaction activities within a span of two months before the customer closes their account. This study managed to group customers into five groups using the k-Means algorithm and then processed using the JRIP algorithm which produced an analysis model and evaluated using 10-fold confusion matrix and cross validation.

Other research, Zorić in 2016 implemented a neural network to predict churn customers in banks using the Alyuda Neuro Intelligence application. This study uses data from 1,866 bank customers in Croatia where the data consists of several information such as gender, age, personal status, average monthly income, internet Funding usage and the use of two or more bank products. The difficulties faced by researchers in this study such as the value of missing data or inconsistent data. For this reason the researcher must directly contact the product owner to obtain the missing data. The results of this study found that there were problematic groups whose contents were young students with less than three product ownership, which in the future could be very important and very valuable customers. In addition, the researchers also found that changing the network topology did not make the results better because all the topologies that were tried produced similar results.

Dolatabadi & Keynia in 2017 tried to use data Extraction models to predict two aspects, namely churn customers and churn employees. The data used is demographic and transaction data. This study uses a comparison of the number of employees or customers who leave with those who remain as much as 15% -20% versus 80-85%. This study uses several data Extraction models, namely Decision Tree, Naïve Bayes, SVM and Neural Network classifiers. From this study, researchers concluded that data Extraction techniques such as SVM can be used in building accurate predictive models to predict churn employees and customers.

Another study conducted by Keramati et al(2016), which is adopted for this study, used three types of data, namely customer dissatisfaction, service use and variables related to customers or customer demographics. This research uses Cross Sectors Standard Process for Data Extraction (CRISP- DM) and uses the Decision Tree model for the modeling phase. There are two reasons researchers use Decision Tree as a prediction model for churn, the first because the Decision Tree can provide easy-to-understand modeling results and the second because of the type of data used. The data used includes numerical and categorical types so that the Decision Tree technique is suitable for this type of data and the results of this study have reached 90% accuracy.

D. Theoretical Framework

Based on the literature study and related works theoretical framework for this research was shown in Figure 5.

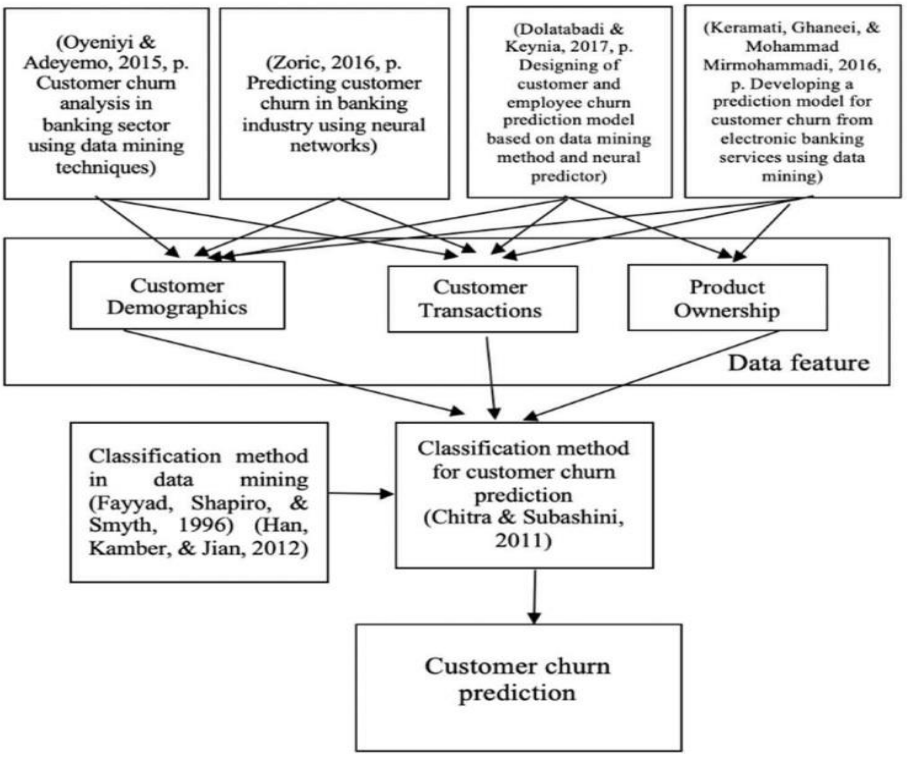

Figure 5: Theoretical Framework

From Figure 5 it can be seen to be able to predict churn customers needed by a data Extraction technique that is catalog technique [7]. The catalog method is a method that is able to predict the class label of an object whose class label is unknown based on training data and test data that has been previously processed [11],[15].

The attributes needed to predict churn customers include customer demographic data [1],[4],[5],[6]. In addition, the catalog method for predicting churn customers is influenced by customer transaction data [1],[4],[5],[6]. The catalog method for predicting churn customers is also influenced by customer product ownership data [4],[5].

This data is used because the tendency of customers to leave can be seen from the number of products they have or the number of products that suddenly change as evidenced by previous research.

III. RESEARCH METHODOLOGY

This research uses deductive method and the type of research is case study research and experimental research. The experiment was conducted by creating a data Extraction learning model that aims to predict customers who will churn. From these problems the research question found is "what is the best catalog model that can be used to predict churn customers, thereby reducing the risk of customers going to Bank XYZ ?". All learning models produced are then evaluated to get the best learning model that best fits the case to be completed. For the research phase, this study uses CRISP-DM as a framework for which the stages of research can be seen in Fig. 6 below.

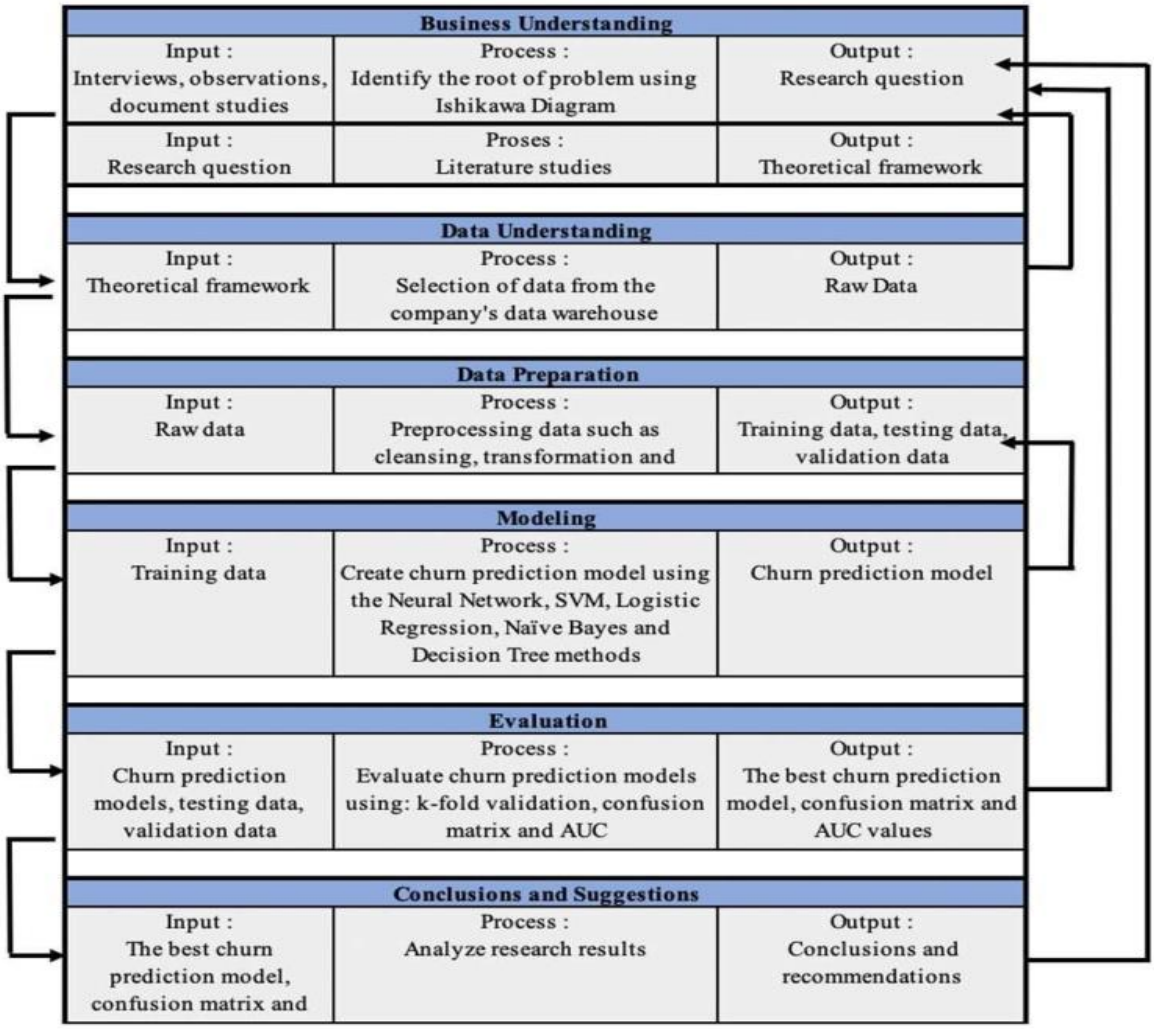

Figure 6: Stages of research using CRISP-DM

At the stage of business understanding, initial data collection is conducted to see problems that occur in the organization. Data collected in the form of market share data published by Bank Indonesia, the number of customers owned by Bank XYZ and the expectations of Bank XYZ concerning the number of these customers. All this information is collected through interviews, observation and document studies. At the stage of understanding data, researchers try to understand what data is needed and available on the XYZ Bank Data Warehouse(D W) . Data selection is based on a theoretical framework that has been made before. The data used are balance data, customer transactions and demography data extracted from the company's Data Warehouse(D W) . Data preparation refers to the formation of dataset training, testing datasets and validation data referring to the preparation of balance data, transactions and demographic data that will be used as input data from the model to be made. In this stage the data transformation and cleaning process is also carried out to support the next modeling process.

Modeling stages to form modeling using a catalog model. The researcher used the RapidMiner application to make modeling. The model used is decision tree, neural network, support vector machine (SVM), naïve bayes and regression logistic. This modeling will produce a customer churn model that is ready to be evaluated. The next stage is to select the model by looking at the results of the validation at the previous stage so that one model will be most suitable for predicting churn customers that occur at Bank XYZ. Finally making conclusions and suggestions is based on the results and analysis of the previous stages.

IV. DATA EXTRACTION PROCESS

A. Data Extraction

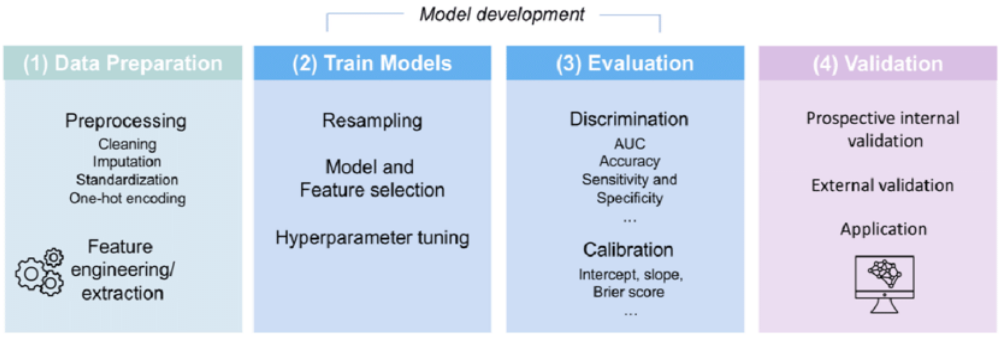

Figure 7: Data extraction

The training and testing data periods used in the study are customer data during 2017 and for validation data using the period January - June 2018. The training and testing data consist of two types of data, namely data from customers closed during the year and customers who remain active until December 2017(see figure 7).

Customer data for training and testing obtained consisted of 57 attributes such as demography (age, vintage, profession, product ownership, etc.), transactions (cash in/out, transactions at owned channel, internal or external transactions, etc.) and balance (average balance in one month, three months, six months, and number of owned accounts) with a total of 131,548 customers divided into 15,949 closed customers and 115,599 active customers. Validation data obtained is 125,707 customers. The number of customers obtained is limited to 1% of the total existing customers by considering the confidentiality of company data and the computational capabilities of the tools used to process learning. Validation data is taken randomly from existing data over a predetermined period.

B. Data Preparation

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 [16].

Data transformation is done on AGE and VINTAGE attributes because the data is still continuous type. To transform it using the Discrete by Entropy function, where the grouping is carried out by the system by defining attributes that were originally numerical or continuous to be nominal attributes. Discretization is done by minimizing entropy so that it can divide data into its best division.

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

C. Catalog Process and Analysis

At this stage a catalog model will be established in accordance with the criteria and settings previously explained. There are five catalog methods that will be used, namely Decision Tree, Neural Network, Support Vector Machine (SVM), Naïve Bayes and Logistic Regression. Comparison of sampling data between classes is divided into three, namely Stratified, 50:50 and 70:30. All models will be validated using k-fold cross validation with the number k = 10. This amount of k = 10 is considered the most ideal number of standards [17].

From the evaluation process, each model will produce true positive, true negative, false positive and false negative values which are then mapped into the confusion matrix to get the performance which can be seen in Table I.

Table 1: Model Performance

METODE	SAMPLING	ACCURACY	RECALL	PRECISION	AUC	VALIDATION TIME (dalam detik)
DECISION TREE	Stratified	91,58%	13,73%	41,05%	0,715	60
	50:50	69,04%	70,58%	16,39%	0,745	59
	30:70	88,47%	39,99%	31,98%	0,759	60
NEURAL NETWORK	Stratified	89,47%	34,61%	34,02%	0,780	122
	50:50	70,79%	69,35%	17,07%	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%**	19,39%	0,811	590
	30:70	89,17%	45,01%	35,65%	0,804	595
NAÏVE BAYES	Stratified	79,65%	60,27%	21,79%	0,793	61
	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

From Table 1 it can be seen if the model formed by sampling 50:50 results in a higher recall value than other sampling measurements. For the model with the highest accuracy, it was produced by SVM with sampling stratified with a percentage of 92.65% but in sensitivity or recall it was still classified as very small with a percentage of 13.92%.

Recall is something important to compare because the recall value is the percentage of success of the model in predicting the true customer churn which is actually the churn of all the customers who actually churn. In other words, this value is the value that represents the success of the company in making the customers who initially want to go to not go. Some models produce recall values that are not too different like SVM with 50:50 sampling or Logistic Regression with 50:50 sampling. The SVM model with 50:50 sampling has a recall percentage of 73.24% and Logistic Regression has a recall of 72.68%.

Precision is the correct percentage of the model predicting the true customer churn which is actually compared to the total of all customer customers predicted by churn. This will have an effect in calculating losses incurred by the company if it follows up on customers who are wrongly predicted. The model that has high precision value is the SVM model with stratified sampling with a percentage of 68.41%. Although this model has a high precision value, the recall value obtained is very low at only 13.92%, so this model cannot be said to be the best model because even though the losses caused by prediction errors are small, the benefits are also very low. The model that has the highest AUC value is Logistic Regression with a sampling of 30:70 but because of its small recall value this model cannot be considered better than the previous two models.

D. Model Selection

The purpose of this study is to solve the problems that exist in the business, so in selecting the model will be mapped into the assumptions of business calculation loss and benefits if this learning model is implemented later. All True Positive predictions are calculated in large amounts assuming the funds have been successfully detained for not leaving the company.

The value of hold able funds is assumed to be the profit obtained by the company if the model is implemented. This is based on the assumption for each customer that is detected by churn and is actually churn after being followed up will still be a customer. The total funds threatened from company data in the validation data were Rp. 624,063,666,740 dues to 9,879 customers. From these data, if each customer is assumed to have the same funds, then the funds threatened to go from each customer is IDR 63,170,733. Therefore, the formula to benefit from each model is True Positive * 63,170,733.

The next assumption is the loss that comes from misdirected follow-up. This misdirected follow-up can be calculated by the number of False Positive times the cost per one follow-up. The cost of a one-time follow-up is assumed to be the same as the one-time cost of the customer making a transaction at the branch of Rp. 15,000 and other marketing costs being ignored. From these assumptions, the formula for calculating losses is False Positive * 15,000. This calculation of loss and profit is made for all the loss and profit calculations that can be seen in Table 2 below.

Table 2: Calculation of Profit and Loss from the Model

METHOD	SAMPLING	TN	FP	FN	TP	Funds Held (in Billions)	Lost (in Billions)	Profit (in Billions)
DECISION TREE	Stratified	112.477	1.947	8.523	1.356	85,660	0,029	85,630
	50:50	78.848	35.576	2.906	6.973	440,490	0,534	439,956
	30:70	106.019	8.405	5.928	3.951	249,588	0,126	249,461
NEURAL NETWORK	Stratified	107.792	6.632	6.460	3.419	215,981	0,099	215,881
	50:50	81.139	33.285	3.028	6.851	432,783	0,499	432,283
	30:70	96.482	17.942	4.880	4.999	315,790	0,269	315,521
SVM	Stratified	113.789	635	8.504	1.375	86,860	0,010	86,850
	50:50	84.348	30.076	2.644	7.235	457,040	0,451	456,589
	30:70	106.396	8.028	5.432	4.447	280,920	0,120	280,800
NAÏVE BAYES	Stratified	93.048	21.376	3.925	5.954	376,119	0,321	375,798
	50:50	88.389	26.035	3.387	6.492	410,104	0,391	409,714
	30:70	91.506	22.918	3.712	6.167	389,574	0,344	389,230
LOGISTIC REGRESSION	Stratified	112.510	1.914	7.801	2.078	131,269	0,029	131,240
	50:50	85.514	28.910	2.699	7.180	453,566	0,434	453,132
	30:70	107.434	6.990	5.876	4.003	252,872	0,105	252,768

From table 2, it can be seen that the calculation of profits is the funds that are saved are reduced by losses obtained from the results of predictions for each model. From the results of these calculations, the SVM model with 50:50 sampling will provide more benefits compared to other models, amounting to 456 billion rupiah. Although the sampling 50:50 Logistic Regression model has a smaller loss, but because the benefits obtained are smaller, it still cannot produce better performance than SVM.

V. CONCLUSION

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

- 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.

- 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 such as Table 2 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. (2017) 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.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

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

[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 Extraction 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 Funding services using data Extraction . Financial Innovation, 2-10.

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

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

[8] De Caigny, A., Coussement, K., & W. De Bock, K. (2018). A new hybrid catalog 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 Extraction 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 Extraction concepts and techniques third edition. Morgan Kaufmann Publishers.

[12] Vafeiadis, T., Diamantaras, K., Sarigiannidis, G., & Chatzisavvas, K. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory 55, 1-9.

[13] Azevedo, A., & Filipe Santos, M. (2008). KDD, semma and CRISP- DM: A parallel overview. European Conference on Data Extraction (pp. 182-185). Amsterdam: IADIS.

[14] Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps. Bangalore: Apress.

[15] Fayyad, U., & Stolorz, P. (1997). Data Extraction and KDD: promise and challenges. Future Generation Computer System 13, 99-115.

[16] Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Catalog with class imbalance problem: A Review. IJASCA Volume 7, 176-204.

[17] Witten, I. H., & Frank, E. (2005). Data Extraction : Practical Machine Learning Tools and Techniques - Second Edition. San Francisco: Morgan Kaufmann Publishers.