Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691| e-ISSN 2964-8246

Customer churn prediction in the case of telecommunication company using support vector machine (SVM) method and oversampling

Dhiya Urrahman[1], Raffi Salman Winanto[2] Thierry Widyatama[3]
[1,2]Department of Informatics Engineering, Universitas Negeri Semarang, Indonesia
[3]Department of Economy and Business, Universitas Dian Nuswantoro, Semarang, Indonesia

Article Info

Article history:

Received November 26, 2024
Revised June 23, 2024
Accepted June 26, 2024

Keywords:

Churn consumers
Machine learning
Support vector machine
Oversampling

ABSTRACT

Churn is the act by which a customer withdraws from service, including service provider-initiated churn and customer-initiated churn. Churn is a big challenge for companies, especially churn-prone enterprise sectors such as telecommunications. Churn can affect both revenue and reputation if occurs for negative reasons. This study aims to predict customer churn in a telecommunication company dataset, investigating the impact of various variables and classes on churn occurrences to inform strategic decision-making for businesses. The Support Vector Machine (SVM) model is employed, and dataset imbalance is addressed through oversampling techniques, specifically Synthetic Minority Over-sampling Technique (SMOTE) and random oversampling (ROS). Three SVM models are created with different training datasets (normal, SMOTE, ROS), yielding varying results. The normal dataset achieves the highest accuracy at 92%, outperforming SVM with ROS (89%) and SVM with SMOTE (87%). However, the normal dataset exhibits lower sensitivity compared to both oversampling techniques. The study identifies the cause of decreased accuracy in oversampling and low sensitivity in the normal dataset. The novelty of this research lies in testing the SVM model's ability to surpass the accuracy of previous models on the same dataset and in exploring the unique impact of oversampling in churn prediction.

[1] ***Corresponding Author:***
Dhiya Urrahman,
Department of Informatics Engineering,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia
Email: urrahmand0@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v2i2.253

1. Introduction

The telecommunications industry has become one of the main sectors that continues to grow in various countries. The increase in technology and the number of service providers has forced many companies to have certain strategies in order to survive in the market. The following are some of the strategies that companies have adopted: (1) trying to get new customers (2) improving the services provided and conducting promotions (3) maintaining subscribed customers over time [1]. Of the three options above, the choice most often taken is to retain customers. Because retaining customers is the most economical way from the existing options and maintaining the credibility of the services provided to consumers [2], [3]. For this reason, many companies are flocking to improve their customer relationship management (CRM) so that customers do not unsubscribe or switch to other providers or service providers [4].

The phenomenon of customer service termination or defection to another provider is called Churn. Churn is a big challenge for many companies, especially churn-prone enterprise sectors such as telecommunications. Churn can have a direct impact on the revenue generated by the company as well as the company's image if churn occurs for negative reasons [5]. Precise forecasting of customer churn can play a pivotal role in formulating strategies for retaining customers and designing cost-effective marketing campaigns. This, in turn, has the potential to result in substantial cost savings for service providers [6]. There are several methods used to predict customer churn, including evaluating the services provided, promotions, customer service, subscription procedures, etc [7]. Telecommunication companies usually have a database that holds the characteristics of their customers, from the type of service, region, length of subscription, and the last review after they churn. This data can be analysed to find out what makes customers churn, so that prediction and prevention can be done. Unfortunately, datasets in companies are usually a collection of unbalanced or imbalanced data [8]. This happens because there are classes that have more or less than other classes in the dataset. However, this can be overcome by dataset pre-processing methods such as sampling which will duplicate the minority of data in the dataset to be trained [9].

The development of technology makes analysing the data collected easier, the most common way companies do is using machine learning to create models that can find out what factors cause churn. In recent times, various classification methods have been employed, particularly in the field of data processing. Researchers commonly favor data mining approaches, utilizing techniques like classification or clustering, such as Decision Trees and K-Means [10]. One of the well-known machine learning models in the classification of a dataset is Support Vector Machine [11]. SVM works by creating a line called a hyperplane in an area that separates many points or data in different classes. In this paper, the SVM method will be used to predict customer churn in telecommunications companies. In its application, Oversampling will be done with Synthetic Minority

Oversampling Technique (SMOTE) and Random Over-Sampling (ROS). SMOTE [12], attempt to balance the class distribution within a dataset by generating synthetic data points for the minority class. SMOTE creates synthetic observations for the minority class in unbalanced data. For each minority class observation, synthetic observations are generated at random between the observation [13]. This method is computationally efficient and thus appropriate for large datasets [14]. The involvement of both techniques will be investigated using the same model and based on findings from another study, SVM can derive advantages from resampling, and various resampling techniques are most effective when paired with specific metrics [15]. The choice of SVM is to minimize the upper bound of generalization error. SVM offers high accuracy, efficient computation time, strong generalization, and overfitting risk reduction capability. As a result, SVM is very suitable for customer churn cases that require high accuracy and time efficiency. The results of the prediction can later be used by companies to consider their decision making.

To stay competitive, companies must adopt new marketing strategies to meet customer needs, increase satisfaction, and retain customers [16]. Churn prediction can be done using supporting data such as churn indicators, customer information (characteristics, subscription details), data usage, billing and payment information, and additional data related to complaints and service improvements. Variables such as income, gender, age, and payment method can also be used in churn prediction models.

2. Method

A flowchart was created that visually described the flow of the method performed, which is presented in Figure 1.

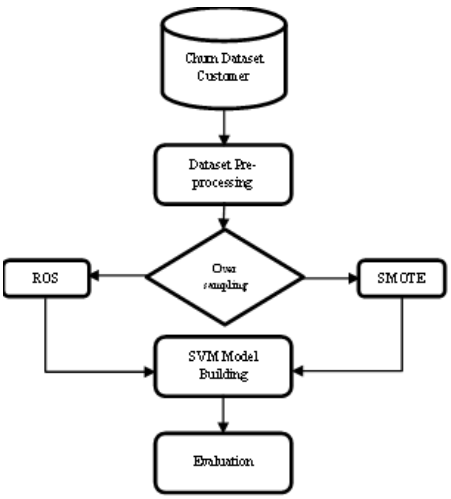

Figure 1. Method flowchart

Customer Churn Dataset

The following section explains the details of the dataset used in the research.

Dataset source

The dataset was obtained from the Kaggle dataset platform which can be accessed at the link https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets. A public dataset from the French telecommunications company called Orange Telecommunications was used, the Orange Telecommunications dataset provides information (features) about user behavior and churn tags that indicate whether a subscription has been canceled. The dataset contains 3000+ records of information about the company's customers and includes up to 20 features.

Dataset characteristics

The dataset has variables about customer characteristics. Figure 2 illustrates the data structure created using RStudio.

```
$ State                 : chr  "KS" "OH" "NJ" "OH" ...
$ Account.length        : int  128 107 137 84 75 118 121 147 141 74 ...
$ Area.code             : int  415 415 415 408 415 510 510 415 415 415 ...
$ International.plan    : chr  "No" "No" "No" "Yes" ...
$ Voice.mail.plan       : chr  "Yes" "Yes" "No" "No" ...
$ Number.vmail.messages : int  25 26 0 0 0 0 24 0 37 0 ...
$ Total.day.minutes     : num  265 162 243 299 167 ...
$ Total.day.calls       : int  110 123 114 71 113 98 88 79 84 127 ...
$ Total.day.charge      : num  45.1 27.5 41.4 50.9 28.3 ...
$ Total.eve.minutes     : num  197.4 195.5 121.2 61.9 148.3 ...
$ Total.eve.calls       : int  99 103 110 88 122 101 108 94 111 148 ...
$ Total.eve.charge      : num  16.78 16.62 10.3 5.26 12.61 ...
$ Total.night.minutes   : num  245 254 163 197 187 ...
$ Total.night.calls     : int  91 103 104 89 121 118 118 96 97 94 ...
$ Total.night.charge    : num  11.01 11.45 7.32 8.86 8.41 ...
$ Total.intl.minutes    : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 11.2 9.1 ...
$ Total.intl.calls      : int  3 3 5 7 3 6 7 6 5 5 ...
$ Total.intl.charge     : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 3.02 2.46 ...
$ Customer.service.calls: int  1 1 0 2 3 0 3 0 0 0 ...
$ Churn                 : chr  "False" "False" "False" "False" ...
```

Figure 2. Dataset Structure

Information that covers the dataset:

- Customers who left in the last month - Churn.
- Services that each customer has signed up for – Internation.plan and Voice.mail.plan.
- Customer account information – Customer.service.calls. Account.length
- Demographic info about the customer – State, Area.code.
- Total customer interactions with the service – Total. dll

Dataset Pre-processing

Before processing, there are several steps taken to optimize the initial dataset obtained from the source in order to improve the performance of the model obtained [17]. The following are some of the methods used in dataset preprocessing.

Determining target variables

The target variables in the training dataset are determined based on the problem statement or analysis objectives. For this case, the target variable to predict is churn.

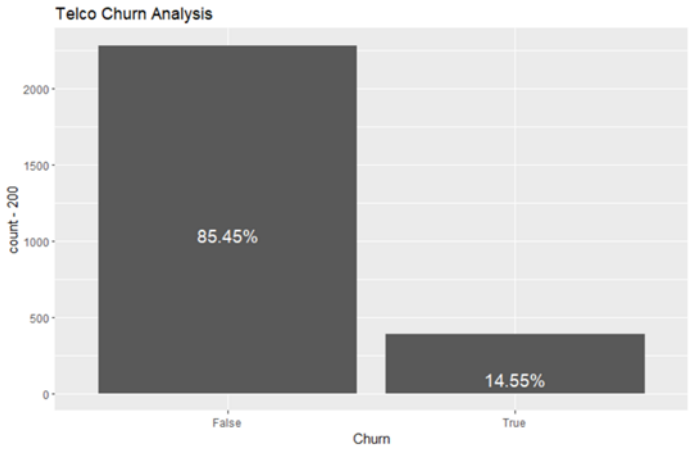

Figure 3. Churn Analysis on Dataset

From Figure 3, it can be seen that customers who churn based on the dataset are 14.55% and those who do not churn are 85.45%.

Dataset partitioning

The research dataset will be subjected to data partitioning. Data partitioning is the division of the dataset into two groups of datasets, namely training datasets and testing datasets with a percentage of 80:20.

Handling Imbalanced Dataset

Dataset imbalance occurs when the distribution of the amount of data in the classes or variables in the dataset is not even or proportional. This means that one class has more entries than another. Real-world datasets frequently encounter class imbalance, where one class comprises fewer instances than the other. Despite being a subject of interest for over two decades, addressing this issue remains a significant and ongoing area of research to enhance accuracy [18]. This happens quite often in customer data, especially churn. because customers who experience churn are far fewer than loyal customers. This makes it difficult to model churn data [19].

In the study of datasets, there is a statistical science that focuses on the selection of data generated from a population of data. This science is called Sampling or commonly known as resampling. Sampling is a typical method used to address the problem of data imbalance in datasets. The level of imbalance can be reduced and classified appropriately using resampling for imbalanced data [20]. Oversampling is a sampling method used to increase the amount of minority class data. Oversampling is done by balancing the distribution of the total amount of data,

usually by automatic duplication, the illustration as in Figure 4. In this study, Random Over-Sampling and SMOTE oversampling techniques were used.

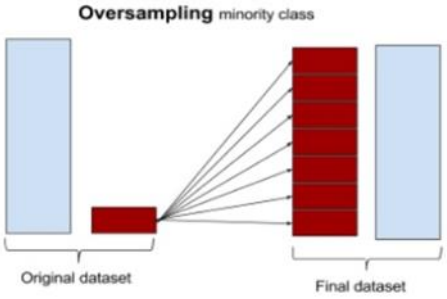

Figure 4. Oversampling Illustration

Random oversampling (ROS)

Random oversampling works by selecting minor class data to be duplicated. However, as the name suggests, the result of random oversampling does not necessarily increase the minority data [21]. Therefore, the minority class does not always become equivalent to the majority class, because if this is done it will be difficult to distinguish data that has features but belongs to different classes if the minority class data continues to be duplicated in large numbers. Random Oversampling (ROS) was chosen in research because of its simplicity and ease of implementation, because it only involves doubling the sample from the minority class until it is balanced with the majority class. This method is effective in dealing with class imbalance without requiring complex parameter adjustments. A simple formula to determine the number of samples that need to be added via Random Oversampling is:

$$N = N_{majority} - N_{minority} \quad (1)$$

Where:

N= The number of samples that need to be added to the minority class.

$N_{majority}$ = The number of samples in the majority class.

$N_{minority}$ = The number of samples in the minority class.

Synthetic minority oversampling technique (SMOTE)

SMOTE works by selecting minority class data and performing K-Nearest Neighbor identification. It creates synthetic data that interpolates between the selected data and its surrounding neighbors, by randomly selecting one of the neighbors and then calculating the difference between the selected data and its neighbors, multiplying it by a random value between 0 and 1 [22]. The result is synthetic data located along the class that requires oversampling, this is repeated continuously until the minority class produces the desired number. Research shows that SMOTE

can improve model performance in terms of metrics such as recall and precision, making them more effective in dealing with class imbalance compared to simple oversampling methods [23], [24].

Oversampling of Training Data

Before Oversampling, the training data has 2278 data for not churn and 388 for churn. SMOTE technique changes the data to 1513 not churn and 1552 churn, while ROS changes the data to 1360 not churn and 1306 churn.

Table 1. Data Balancing

Technique	Program	Churn	
		No	Yes
Normal	The training dataset has been split from the source	2278 (85%)	388 (15%)
SMOTE	> train_oversampled <- SMOTE(Churn ~ ., smotetrain_df, k = 5, perc.over = 300, under = TRUE, perc.under = 135)	1571 (51%)	1552 (49%)
ROS	> train_oversampled <- ROSE(Churn ~ ., data = train_df, seed = 2021)$data	1360 (51%)	1306 (49%)

It can be seen from Table 1 that the results of data balancing using oversampling are equivalent for both methods. However, in program usage, SMOTE has more configuration options than ROS.

Support Vector Machine (SVM) Modeling

Support Vector Machine (SVM) is a supervised learning technique that can generate predictions using a linear combination of kernel basis functions for regression and classification problems. Vladimir Vapnik and his colleagues at AT&T Bell Laboratories introduced and developed SVM. Instead of limiting empirical error, SVM implements the principle of structure risk minimization (SRM) by minimizing the upper bound of generalization error [11]. The attenuation of the input data to a high-dimensional space, where the data will be linearly divided, is used with a kernel in the classification problem. Kernels are used to find a hyperplane that can minimize the distance between two datasets for regression problems.

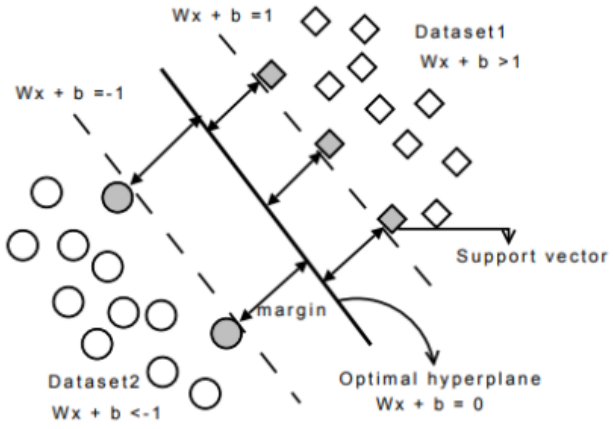

Figure 5. SVM Illustration

The main purpose of SVM is to find the best hyperplane in N-dimensional space, SVM is used. In Figure 5, we can see the best hyperplane displayed in SVM using two large datasets present in N-dimensional space. Support Vectors (SVs) are vectors that are located as close to the hyperplane as possible. Since this parameter has a significant influence on how well the kernel method performs, the output of the SVM model depends on its use. Many parameters are at the margin of outputting a number of datasets.

In Support Vector Machine modeling, modeling is done using a linear kernel as in Table 2. The results obtained from the modeling are evaluated for classification using testing data.

Table 2. Support vector machine modeling

Technique	Program	Number of Support Vectors
Normal	> fit.svm <- svm(Churn~., data=train_df)	772
SMOTE	> smotefit.svm <- svm(Churn~., data=smotetrain_df)	1408
ROS	> rosefit.svm <- svm(Churn~., data=rosetrain_df)	1482

In SVM (Support Vector Machine), "Number of Support Vectors" refers to the number of data points from the training set that are identified as support vectors. Support vectors are data points that are closest to the decision boundary or have significant influence in determining the decision boundary in SVM. When training an SVM model, the algorithm aims to find the optimal hyperplane that separates the different classes in the training data. Support vectors are data points that lie on or near the decision boundary or misclassification boundary. These support vectors play an important role in determining the decision boundary and determining the classification boundary of the SVM model. Having more support vectors can indicate more complex decision boundaries, which can result in a more flexible and potentially overfitting model [25]. Therefore, it is important to strike a balance

between the number of support vectors and model complexity to achieve good generalization performance on unseen data.

Evaluation

A performance function is created to calculate the specificity and accuracy of the model used based on the confusion matrix or contingency table in the model [26]. Then applied SVM classification to fit.svm to test and generalize the prediction label for each sample in the dataset.

```
# performance function
performance <- function(table, n=2){
  tn = table[1,1]
  fp = table[1,2]
  fn = table[2,1]
  tp = table[2,2]

  sensitivity = tp/(tp+fn) # recall
  specificity = tn/(tn+fp)
  ppp = tp/(tp+fp) # precision
  npp = tn/(tn+fn)
  hitrate = (tp+tn)/(tp+tn+fp+fn) # accuracy

  result <- paste("Sensitivity = ", round(sensitivity, n) ,
                  "\nSpecificity = ", round(specificity, n),
                  "\nPositive Predictive Value = ", round(ppp, n),
                  "\nNegative Predictive Value = ", round(npp, n),
                  "\nAccuracy = ", round(hitrate, n), "\n", sep="")

  cat(result)
}
```

Figure 6. Labeling code

The program will produce a confusion matrix or confusion table that contains the prediction results of the model created.

Table 3. Prediction Result

Technique	Observation	Prediction	
		Not Churn	Churn
Normal	Not Churn	565	7
	Churn	45	50
SMOTE	Not Churn	512	60
	Churn	30	65
ROS	Not Churn	527	45
	Churn	26	69

Based on Table 3, the SVM classifier makes the following predictions on the test dataset:

- Sample predicted negative (No) and actually negative (true negative, TN).
- Samples predicted negative (No) but actually positive (false negative, FN).
- Sample predicted positive (Yes) but actually negative (false positive, FP).
- Sample predicted positive (Yes) and actually positive (true positive, TN).

This value can be used to calculate various performance metrics for the classifier, such as accuracy, precision, and sensitivity. Here is an example of how to use the previously defined performance( ) function to calculate the metrics.

Table 4. Performance measures

Technique	Performance				
	Sensitivity	Spesificity	Predicted Positive Value	Predicted Negative Value	Accuracy
Normal	0.53	0.99	0.88	0.93	0.92
SMOTE	0.68	0.9	0.52	0.94	0.87
ROS	0.73	0.92	0.61	0.95	0.89

Table 4 show that the classification model has:

- Sensitivity means the percentage of correctly identified positive cases
- Specificity means the percentage of correctly identifying negative cases.
- Positive predictive value means the percentage of cases predicted by the model to be positive are actually positive.
- Negative predictive value means that the percentage of cases predicted by the positive model are actually negative.
- Accuracy means the percentage of all cases correctly classified by the positive model.

3. Results and Discussion

Analysis of Research Results

Based on the results obtained from the research conducted, churn prediction using SVM provides the highest accuracy of 92% using a normal dataset without oversampling. These results were obtained with a little manual feature selection performed on the original dataset, namely eliminating the State, Account.length, and Area.code features. Due to the imbalance of the dataset, research is also proposed using several oversampling methods.

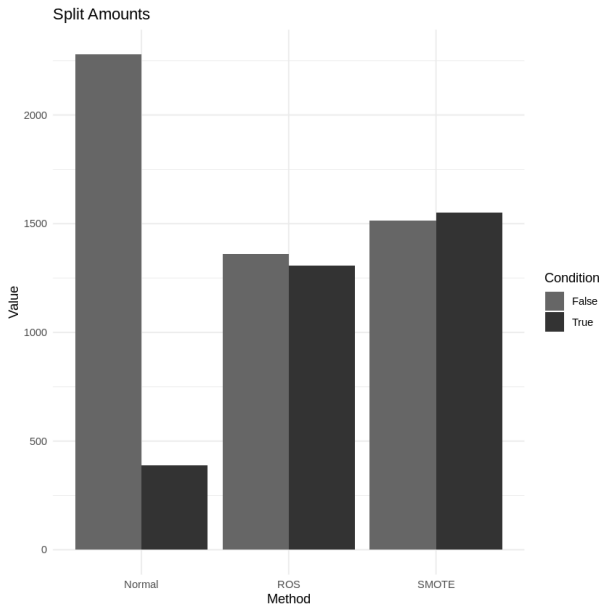

Figure 7. Dataset Balance Chart

It can be seen in Figure 7 that, oversampling successfully balances the percentage of churn and non-churn data on the dataset.

Model Evaluation

The accuracy obtained from the oversampled training dataset cannot pass the accuracy of the normal dataset, where SMOTE can only achieve 87% accuracy while ROS reaches 89%. Accuracy comparison can be seen in the Table 5.

Table 5. Accuracy results

	Method		
	Normal	SMOTE	ROS
Accuracy	92%	87%	89%

This happens because of several possibilities, including:

Overfitting: Synthesized data derived from oversampling is likely to duplicate minority data containing noise. Learning too well on such data leads to poor generalization of the model which causes a decrease in model performance on testing data [27].

Data Quality: The oversampling technique generates synthesized data based on the minority class or variable in the dataset. If the original data in the minority contains noise, is ambiguous or contains errors, then the duplicated data also has the same characteristics [28]. As a result, oversampling data contains more inaccuracies than normal data and results in lower accuracy.

Looking at these two conjectures, it appears that oversampling seems to be an obstacle to model performance. For this reason, further analysis was conducted on the performance of the three models which resulted in the following results

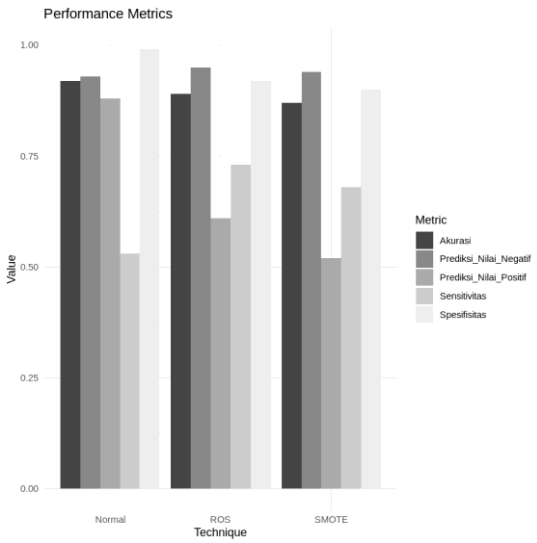

Figure 8. Model Performance Graphic Chart.

Figure 8 shows the superiority of the normal model over the oversampling model in terms of Accuracy, Predicted Positive Value and Specificity, while the sensitivity metrics lag far behind the two oversampling models. This is due to the difference in dataset balance discussed earlier. The lagging minority class in the normal dataset causes difficulty in the learning model to identify positive values (not churn) that are actually positive [20]. The imbalance creates a bias towards the majority class in the training data which causes a tendency for accuracy to favor the majority data. The reduction of majority data in oversampling models increases the sensitivity of the model which allows the model to learn more discriminative decision boundaries, making it more sensitive in capturing true positive test data and reducing false negatives [28]. These two things make the oversampling model have an advantage in Negative Value Prediction and Sensitivity over the normal model. Therefore, the use of oversampling is still very feasible to be done in churn prediction which has imbalance, especially for sensitivity improvement. It's just that further addressing is needed for minority data classes and variables that will be oversampled so that overfitting does not occur due to errors and noise in minority data [27].

Comparison with Previous Research

The proposal of the SVM model and oversampling in this study is based on research by Lawchak Fadhil Khalid et al. in 2021 regarding Customer Churn Prediction in the Telecommunications Industry Based on Data Mining [29]. In this study, data mining procedures were carried out on the dataset used to obtain the optimal feature selection in pre-processing the dataset, then several models were made to predict churn. The models used include Decision Tree with 94% accuracy, Random Forest with 91% accuracy, Neural Network with 90% accuracy, and Naïve Bayes with 88% accuracy. Seeing these results, new research is proposed using the same

dataset, a model that has not been made, namely SVM and the use of oversampling to replace data mining. The comparison obtained is that the SVM model using normal datasets managed to surpass 3 models from the four models tested in the reference research with an accuracy rate of 92% and the use of ROS oversampling managed to rival the use of the Naïve Bayes model with 89% accuracy.

4. Conclusion

Customer churn prediction research in telecommunication companies is conducted to identify variables or classes that affect the occurrence of churn by customers. The Support Vector Machine (SVM) model was used by considering its superiority in predicting churn and obtained an accuracy of 92%. Seeing that the unbalanced dataset allows the use of oversampling on minority classes in the training dataset, the Random Over-Sampling and SMOTE methods are used. However, the results obtained using oversampling actually produce lower accuracy, namely ROS with 89% and SMOTE with 87%. Further analysis and comparison of the three models were conducted and it was concluded that the decrease in the oversampling method was due to errors and noise in the duplicated minority class which caused a decrease in model performance when compared to normal training data. Looking at the performance metrics created, it is seen that the normal dataset lags behind in terms of sensitivity and prediction of negative values, this is due to the bias in the majority class which creates a weakness in the model to detect genuine positives in the test data. Looking at the evaluation, there are advantages in using oversampling in churn data prediction and factors that need to be considered in its use. With the results obtained in this study, companies can perform churn prediction with a tested model and choose the pre-processing stage of the dataset needed depending on the desired results and capabilities of the model.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Thierry Widyatama, the third author, for generously dedicating his time to review the article we have prepared. His valuable insights and thoughtful feedback have significantly contributed to the refinement of our work. We extend our appreciation for his unwavering support and commitment to ensuring the quality of this article. Thierry Widyatama's expertise and dedication have been instrumental in enhancing the overall depth and clarity of our research. Additionally, we would like to acknowledge and thank any sponsors or financial supporters who have played a role in making this research possible. Thank you once again to Thierry Widyatama and all those who have contributed to the success of this project.

REFERENCES

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[2] Y. Pandey, R. Jha, and U. Umamaheswari, "CUSTOMER CHURN ANALYSIS IN TELECOM ORGANIZATION," *J. Posit. Sch. Psychol.*, pp. 5475–5488, 2022.

[3] A. Nazal and Y. Megdadi, "The Role of Customer Relationship Management Strategies on Developing Customer Services of Jordanian Telecommunication Companies," *J. Mark. Manag.*, vol. 7, pp. 77–88, Dec. 2019, doi: 10.15640/jmm.v7n2a9.

[4] L. Geiler, S. Affeldt, and M. Nadif, "A survey on machine learning methods for churn prediction," *Int. J. Data Sci. Anal.*, vol. 14, no. 3, pp. 217–242, Sep. 2022, doi: 10.1007/s41060-022-00312-5.

[5] Y. Zhang, S. He, S. Li, and J. Chen, "Intra-Operator Customer Churn in Telecommunications: A Systematic Perspective," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 948–957, Jan. 2020, doi: 10.1109/TVT.2019.2953605.

[6] H. Faris, "A Hybrid Swarm Intelligent Neural Network Model for Customer Churn Prediction and Identifying the Influencing Factors," *Information*, vol. 9, no. 11, p. 288, Nov. 2018, doi: 10.3390/info9110288.

[7] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Investigating factors affecting customer churn in electronic banking and developing solutions for retention," *Int. J. Electron. Bank.*, vol. 2, no. 3, p. 185, 2020, doi: 10.1504/IJEBANK.2020.111427.

[8] A. Viloria, O. B. Pineda Lezama, and N. Mercado-Caruzo, "Unbalanced data processing using oversampling: Machine Learning," *Procedia Comput. Sci.*, vol. 175, pp. 108–113, 2020, doi: 10.1016/j.procs.2020.07.018.

[9] P. Joshi and S. Gupta, "Predicting Customers Churn in Telecom Industry using Centroid Oversampling method and KNN classifier," *Int. Res. J. Eng. Technol.*, vol. 6, no. 4, pp. 3708–3712, 2019.

[10] S. Arifin and F. Samopa, "Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method," *J. Phys. Conf. Ser.*, vol. 1108, p. 012018, Nov. 2018, doi: 10.1088/1742-6596/1108/1/012018.

[11] X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 2, pp. 458–475, Apr. 2022, doi: 10.3390/jtaer17020024.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[13] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100222, Jun. 2024, doi: 10.1016/j.caeai.2024.100222.

[14] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[15] B. Zhu, B. Baesens, A. Backiel, and S. K. L. M. vanden Broucke, "Benchmarking sampling techniques for imbalance learning in churn prediction," *J. Oper. Res. Soc.*, vol. 69, no. 1, pp. 49–65, Jan. 2018, doi: 10.1057/s41274-016-0176-1.

[16] S. Baker, B. Baugh, and M. Sammon, "Measuring Customer Churn and Interconnectedness," Cambridge, MA, Aug. 2020. doi: 10.3386/w27707.

[17] S. Pandya and P. Mehta, *A Review On Sentiment Analysis Methodologies, Practices And Applications*. 2020.

[18] S. Sharma, A. Gosain, and S. Jain, "A Review of the Oversampling Techniques in Class Imbalance Problem," 2022, pp. 459–472. doi: 10.1007/978-981-16-2594-7_38.

[19] I. V. Pustokhina, D. A. Pustokhin, P. T. Nguyen, M. Elhoseny, and K. Shankar, "Multi-objective

rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector," *Complex Intell. Syst.*, vol. 9, no. 4, pp. 3473–3485, Aug. 2023, doi: 10.1007/s40747-021-00353-6.

[20] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New imbalanced bearing fault diagnosis method based on Sample-characteristic Oversampling TechniquE (SCOTE) and multi-class LS-SVM," *Appl. Soft Comput.*, vol. 101, p. 107043, Mar. 2021, doi: 10.1016/j.asoc.2020.107043.

[21] M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction In Banking," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2020, pp. 1196–1201. doi: 10.1109/ICECA49313.2020.9297529.

[22] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/joscex.v1i1.5.

[23] S. J. Haddadi, A. Farshidvard, F. dos S. Silva, J. C. dos Reis, and M. da Silva Reis, "Customer churn prediction in imbalanced datasets with resampling methods: A comparative study," *Expert Syst. Appl.*, vol. 246, p. 123086, Jul. 2024, doi: 10.1016/j.eswa.2023.123086.

[24] C. Rao, Y. Xu, X. Xiao, F. Hu, and M. Goh, "Imbalanced customer churn classification using a new multi-strategy collaborative processing method," *Expert Syst. Appl.*, vol. 247, p. 123251, Aug. 2024, doi: 10.1016/j.eswa.2024.123251.

[25] S. Ougiaroglou, K. I. Diamantaras, and G. Evangelidis, "Exploring the effect of data reduction on Neural Network and Support Vector Machine classification," *Neurocomputing*, vol. 280, pp. 101–110, Mar. 2018, doi: 10.1016/j.neucom.2017.08.076.

[26] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

[27] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. A. M. Santos, "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]," *IEEE Comput. Intell. Mag.*, vol. 13, pp. 59–76, 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:52986310

[28] Y. Kim, Y. Kwon, and M. C. Paik, "Valid oversampling schemes to handle imbalance," *Pattern Recognit. Lett.*, vol. 125, pp. 661–667, Jul. 2019, doi: 10.1016/j.patrec.2019.07.006.

[29] N. Mustafa, L. Sook Ling, and S. F. Abdul Razak, "Customer churn prediction for telecommunication industry: A Malaysian Case Study," *F1000Research*, vol. 10, p. 1274, Dec. 2021, doi: 10.12688/f1000research.73597.1.