Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of Random Forest Algortihm on Sales Data to Predict Potential Churn of Products

Abdi Dharma[1)*], Christnatalis[2)], Windy Candra[3)], Josua Presen Turnip[4)]
[1,2,3,4)] Universitas Prima Indonesia, Indonesia
[1)] abdidharma@unprimdn.ac.id, [2)]christnatalis@unprimdn.ac.id, [3)]windycandraa@gmail.com, [4)]josuaturnip10@gmail.com

Submitted : Mar 31, 2023 | **Accepted** : Apr 1, 2023 | **Published** : Apr 1, 2023

Abstract: Concentration of sales that are focused on products that are in great demand and are popular is one of the supermarket sales techniques. Seasonal sales techniques like this sometimes have an impact that can be seen obviously by the imbalance in sales of existing products in supermarkets. Sales imbalance can be the initial cause for a product to lose interest and become a product that is eventually removed from store. With a classification model made to predict which products will be eliminated or churn, it can assist staff in distributing the sales of each product. The more products are churn due to lack of enthusiasts which can affect the overall sales of the supermarket. The lower the sales, the slower the development of supermarket to develop. The purpose of this study is to assist staff in classifying potentially churn products. The information gain from classification task can assist the marketing staff to determine sales strategy to improve sales on potential churn products .The classification model consists of 3 models with different algorithms and the results show that the application of the Random Forest algorithm is more effective for predicting data with 96% accuracy compared to 81% for the Logistic Regression algorithm and 46% for the Support Vector Machine algorithm.

Keywords: Churn; Synthetic minority over-sampling technique (SMOTE); *Logistic Regression; Random Forest; Support Vector Machine*

INTRODUCTION

This study focuses on predicting potential churn products of supermarket using Logistic Regression, Random Forest, and the Support Vector Machine algorithm. Starting with a problem of any supermarket, the imbalance sales distribution may affect the development of the supermarket. The imbalance sales of any product are the initial factor of product reduction. With fewer option of products and less stock on market, this will result in delays on development of supermarket and the market share among supermarkets in the region (Dingli et al., 2017). Product Churn Prediction itself is a process for determining strategies to provide sales improvement by marketing staff. Implementation of sales stabilization includes predicting and preventing product to churn (Wardani N et al., 2018). Application of classification model to predict churn on product is to help marketing staff to prepare strategy to prevent product to churn (Kartika Sari & Wahyu Wibowo, 2018). With the help of classification model, it is expected to assist in delaying a decline in sales by making a decision before product is churn.

Then the latest contribution to this research is to avoid products to become less as it is eliminated. Previously, churn prediction was generally carried out on datasets from telecommunication industry

*abdidharma@unprimdn.ac.id

using Logistic Regression, Random Forest, and Decision Tree (Wicaksono, 2021). This research will examine the churn of customers in the telecommunication industry, but this research focuses specifically on predicting churn on supermarket products. Which algorithm is most suitable to classify the potential churn products is the purpose of the research.

LITERATURE REVIEW

Various algorithms that have been studied include Logistic Regression, Random Forest, and Decision Tree. The study of UJI PERFORMA TEKNIK KLASIFIKASI UNTUK MEMPREDIKSI CUSTOMER CHURN shows that Logistic Regression algorithm is most suited to predict customer churn in the telecommunication industry to prevent loss income from the churn customers. This research itself is using the Logistic Regression algorithm because this algorithm when combine with Backward Elimination has been success to use for classification task with 82.23% accuracy, 57.22% recall, and 0.853 AUC. In this study, three different algorithms have been applied to get the best algorithm to perform the classification task to predict potential churn by using Linear Regression, Decision Tree, and Random Forest Algorithm. Logistic Regression algorithm is an advance development of Linear Regression algorithm. Linear Regression only handles one independent variable and one dependent variable where Logistic Regression handle more than one independent variables. Random Forest algorithm is development of Decision Tree algorithm when Decision Tree just find one optimal node of target, Random Forest combines more than one Trees to find the optimal result. Support Vector Machine classify data by dividing data class by hyperplane margin line. Logistic Regression, Random Forest, and SVM are widely used to classify data such as age, gender, temperature, types, etc.

METHOD

This study will conduct an experimental analysis using collected sales data from Suzuya Supermarket on PT. Suriatama Mitra Perwita. Then, the data is processed with three different algorithm Logistic Regression, Random Forest, and Support Vector Machine to classify the active status of products on supermarket.

(1) Data Collection
The dataset used in this study was obtained from Suzuya Supermarket on PT. Mitra Perwita. This research uses sql query to collect the data in January 2022 with a total of 316 products sales.

(2) Logistic Regression
Logistic Regression is one of the generalized linear models (GLM). It is a model for binary variable where the response records either success or failure for a given event (Jawa, 2022). Logistic regression can be extended to combine more than one independent variable, which can be continuous or categorical variables (Fosdal, 2017).

(3) Random Forest
Random forest is an ensemble learning algorithm based on decision trees(Irmanda et al., 2019) . It uses bootstrap technology to extract randomized samples from original samples to construct a single decision tree (Zhu et al., 2021). At each node of the decision tree, a random feature subspace is used to select a sorting point. Random Forest is developed algorithm of decision tree where is the simple clasiffier model (Gu et al., 2021).

(4) Support Vector Machine
Support Vector Machines (SVM) are a supervised learning algorithm, first developed and published by Vapnik and Lerner (1963) (Ay, Stemmler, et al., 2019). Vapnik (1995) later extended the theory of SVM to regression problems. SVM can just be defined as a prediction tool wherein we search for a particular line or decision boundary termed as hyperplane which easily separates out the datasets or classes, hence it avoids the extra over fit to the data (Ay, Stenger, et al., 2019).

*abdidharma@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

(5) Flow of Research Methods

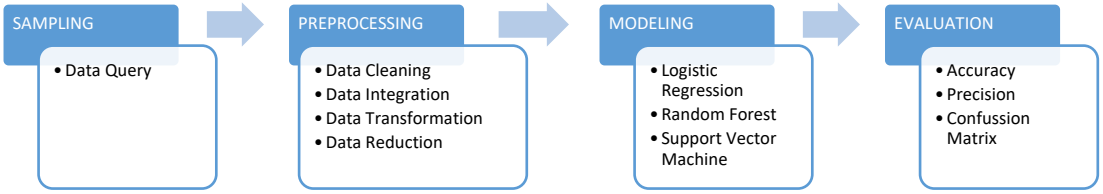

Figure 1. Flowchart of Research Methods

The flow of the research method is carried out according to the flowchart as shown in the image above, including:

a) Data Collection: In this stage, dataset is obtained by query data from Suzuya Supermarket database.
b) Preprocessing: After the data is collected, the data will be cleaned, and transformed to reduce the redundant data.
c) Modelling: Data that has been preprocessed is then divided into training data and testing data to process with 3 different classification algorithms to ensure accurate analysis results.
d) Evaluating: Result data will be analyzed by comparing the quality of accuration of each algorithm

RESULT

Data Sales

Table 1. Top 3 Rows of Sales Dataframe

Row	Product	Description	Size	UOM	Min	Max	EndQ	202112 SQty	202111 SQty	202110 SQty	202109 SQty
1	499712	PANTENE CONDITIONER SMOOTH&SILKY 135ML	135ML	PCS	13	22	53	13	14	23	19
2	499713	PANTENE CONDITIONER SMOOTH&SILKY 75ML	75ML	PCS	0	0	0	0	0	0	0
3	499813	PANTENE SHAMPOO LONG BLACK 290ML	290ML	PCS	26	43	60	39	27	34	24

The sales data are collected using sql query that help to obtained data from database. In this stage, the collected data are focus on Suzuya Supermarket. After retrieving the data, all the information obtained such as product details, size, unit of measurement, end quantity, and monthly sales data are collected and stored in the form of a CSV (Comma Separated Value) file to be entered into the Jupyter Notebook program in the Dataframe format as shown above.

*abdidharma@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Preprocessing

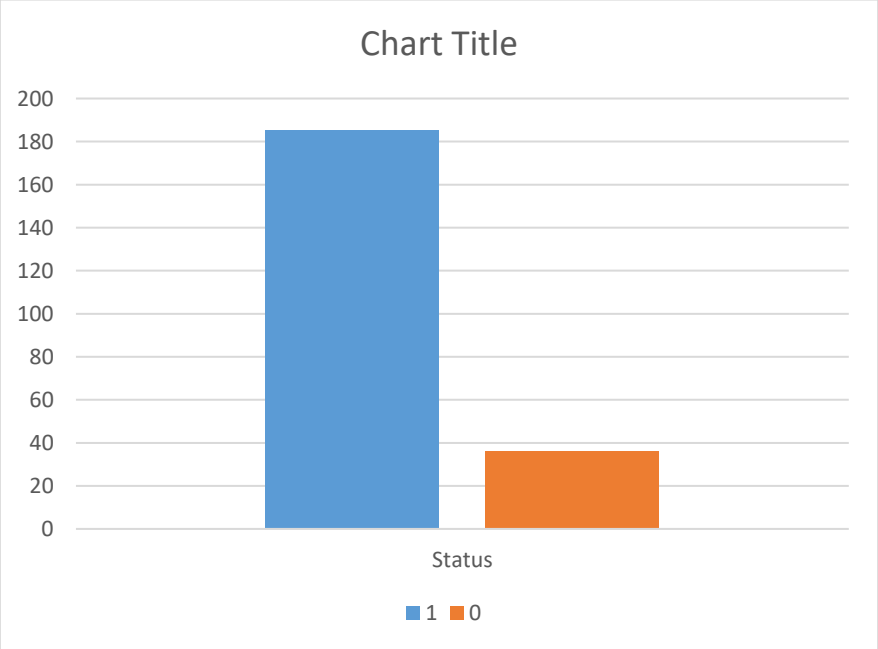

Figure 3. Target Feature Before SMOTE

In this stage, the target data are found imbalance that will affect the quality of dataframe. In some case when imbalanced ratio are high, samples that caused the lack of data will be removed (Fernández et al., 2018). SMOTE (Synthetic Minority Over-Sampling Technique) is some of popular methode to handle imbalance class (Dablain et al., 2022). Imbalance data caused by amount of object class are more than another class. The method is done by generate synthetic sample from minority class by create combination convex from nearest instance to balancing data (Azmatul Barro et al., 2013).

Table 2. Target Feature Before and After SMOTE Process

Status 0 Before SMOTE	Status 1 Before SMOTE	Status 0 After SMOTE	Status 1 After SMOTE
36	185	185	185

Data Split

In this study, the data used is divided into two different parts, namely training data and testing data. Spliting data is divided by 70% of train data and 30% of test data. This aims to ensure that the built model can classify new data. Train data is used to build the model and test data is used to evaluate model performance. By sharing this data, we can ensure that our model performs well and is reliable in real situations.

Prediction Results with Logistic Regression

Prediction Results with Logistic Regression algorithm in the process of predicting products status in this study. Logistic Regression algorithm predict the data by calculated variables into linear equation on sigmoid function. The performance results from the testing process show that Logistic Regression model has an accuracy of 81%. The results of this performance illustrate that the Logistic Regression model has a fairly good performance in predicting data, but it is not satisfactory.

*abdidharma@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3. Logistic Regression Confusion Matrix

		Predicted Class	
		1	0
Actual Class	1	77	0
	0	18	0

Prediction Results with Random Forest

Prediction Results with Random Forest algorithm in the process of predicting products status with multiple decision trees. Decision Tree is a basic classification method that starts from the root node of the tree and continues to divide by selecting the optimal attribute to form nodes one by one until the goal condition is reached. The performance results from the testing process show that Random Forest model has an accuracy of 96%. The results of this performance illustrate that the Random Forest model has a greate performance in predicting data.

Table 4. Random Forest Confusion Matrix

		Predicted Class	
		1	0
Actual Class	1	76	1
	0	2	16

Prediction Results with Support Vector Machine

Prediction Results with Support Vector Machine algorithm in the process of predicting products status with help of margin line called “hyperplane” that divide class by the nearest instance (Somvanshi & Chavan, 2016). The performance results from the testing process show that Support Vector Machine model has an accuracy of 46%. The results of this performance illustrate that the Support Vector Machine model is not good enough in performance to predicting the existing data.

Table 5. Support Vector Machine Confusion Matrix

		Predicted Class	
		1	0
Actual Class	1	29	48
	0	3	15

*abdidharma@unprimdn.ac.id

The Importance of Sales Data Classification Analysis

Sales data classification is very important because it provides information about the prediction of potential churn product data. This is very useful for the marketing and merchandise staff to evaluate sales and decide programs to help the potentially churn data are not eliminated from store.

Comparison of Algorithms

Based on the evaluation results of the built models, it can be said that the Random Forest model has better performance than the Logistic Regression and Support Vector Machine model in terms of predicting product churn at supermarket. This can be seen from the results of training and validation accuracy in each model. The Random Forest model has an accuracy of 96% while the Logistic Regression model only has an accuracy of 81% and the Support Vector Machine only has an accuracy of 46%. Another advantage of the Random Forest model is that it classify data with multiple decision tree classifier.

DISCUSSIONS

Based on the evaluation results of the built models, with the accuracy results of different models as 96% of Random Forest model, 81% of Logistic Regression, and 46% of Support Vector Machine, it can be said that the Random Forest model has better performance. It shows that Random Forest algorithm can provide more accuracy by using repetitive of decision trees to find the optimal node of target. Logistic Regression has lower performance to classify the data by model the relation between independent variables and dependent variable than Random Forest model.

CONCLUSION

Analysis of sales data predictions on the Suzuya Supermarket obtained is provides accurate and useful information for marketing staff. Hotel review predictions can help marketing staff to find out products are in loss sales and potentially churn. This information can help marketing satff to determine corrective actions to provide sales to predicted products. In this study, the Logistic Regression, Random Forest and Support Vector Machine models have been used to predict potentially churn products. The results of the analysis show that the Logistic Regression model has a fairly good performance but is still less effective in predicting churn products with an accuracy of 81%. Meanwhile, the Random Forest model has better performance with an accuracy of 96% and Support Vector Machine with the lowest result only 46% of accuracy. This shows that Random Forest algorithm has the potential to improve the effectiveness of predicting churn products by monthly sales report.

REFERENCES

Ay, M., Stemmler, S., Schwenzer, M., Abel, D., & Bergs, T. (2019). Model predictive control in milling based on support vector machines. *IFAC-PapersOnLine*, *52*(13), 1797–1802. https://doi.org/10.1016/j.ifacol.2019.11.462

Ay, M., Stenger, D., Schwenzer, M., Abel, D., & Bergs, T. (2019). Kernel Selection for Support Vector Machines for System Identification of a CNC Machining Center. *IFAC-PapersOnLine*, *52*(29), 192–198. https://doi.org/10.1016/j.ifacol.2019.12.643

Azmatul Barro, R., Sulvianti, I. D., & Afendi, M. (2013). *PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) TERHADAP DATA TIDAK SEIMBANG PADA PEMBUATAN MODEL KOMPOSISI JAMU* (Vol. 1, Issue 1).

Dablain, D., Krawczyk, B., & Chawla, N. v. (2022). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2021.3136503

Dingli, A., Marmara, V., & Fournier, N. S. (2017). Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International Journal of Machine Learning and Computing*, *7*(5), 128–132. https://doi.org/10.18178/ijmlc.2017.7.5.634

Fernández, A., García, S., Herrera, F., & Chawla, N. v. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. In *Journal of Artificial Intelligence Research* (Vol. 61).

*abdidharma@unprimdn.ac.id

Fosdal, S. (2017). *The use of logistic regression and quantile regression in medical statistics*.

Gu, H., Yang, M., Gu, C. shi, & Huang, X. fei. (2021). A factor mining model with optimized random forest for concrete dam deformation monitoring. *Water Science and Engineering*, *14*(4), 330–336. https://doi.org/10.1016/j.wse.2021.10.004

Irmanda, H. N., Astriratma, R., & Afrizal, S. (2019). PERBANDINGAN METODE JARINGAN SYARAF TIRUAN DAN POHON KEPUTUSAN UNTUK PREDIKSI CHURN Universitas Pembangunan Nasional Veteran Jakarta. *JSI : Jurnal Sistem Informasi (E-Journal)*, *11*(2). http://ejournal.unsri.ac.id/index.php/jsi/index

Jawa, T. M. (2022). Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia. *Alexandria Engineering Journal*, *61*(10), 7995–8005. https://doi.org/10.1016/j.aej.2022.01.047

Kartika Sari, Y., & Wahyu Wibowo, F. (2018). Prediksi Customer Churn Berbasis Adaptive Neuro Fuzzy Inference System. In *Generation Journal* (Vol. 2, Issue 1). www.internetworldstats.com,

Somvanshi, M., & Chavan, P. (2016). *A Review of Machine Learning Techniques using Decision Tree and Support Vector Machine*.

Wardani N, Dantes G, & Indrawan G. (2018). *PREDIKSI CUSTOMER CHURN DENGAN ALGORITMA DECISION TREE C4.5*.

Wicaksono, A. (2021). *Uji Performa Teknik Klasifikasi untuk Memprediksi Customer Churn*. *9*(1).

Zhu, L., Zhou, X., & Zhang, C. (2021). Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. *Artificial Intelligence in Geosciences*, *2*, 76–81. https://doi.org/10.1016/j.aiig.2021.12.001

*abdidharma@unprimdn.ac.id