Transaction on Informatics and Data Science
Vol. 1, No. 2, 2024, 64-72
ISSN xxxx-xxxx (online); xxxx-xxxx (print)
DOI: https://doi.org/10.24090/tids.v1i2.12277

Prediction of Customer Switching Using Support Vector Machine Method

Abu Tholib *[1], Selfia Hafidatus Sholeha[1], Qurrotu Aini[1]

[1] Informatics Engineering, Faculty of Engineering, Nurul Jadid University, Paiton, Probolinggo, Indonesia, ebuenje@gmail.com

Article Information

Submitted March 12, 2024
Accepted May 28, 2024
Published October 1, 2024

Abstract

Several studies on predicting customer switching focus on the telecommunications industry and online stores. This research aims to predict customer switching to get the best results; customers are the most critical mass; some companies must provide satisfying services so customer flow decreases. The support vector machine (SVM) method uses machine learning to find a hyperplane based on the SRM principle. A hyperplane is a decision boundary that helps classify data points. SVM stands out for its ability to take input data and make predictions based on its characteristics. This study uses data from Kaggle, structured it, cleaned it, identified patterns and inconsistencies (such as skewness, outliers, and missing values), and built and validated hypotheses. From the data processing, the plot shows the imbalance of data classes between churners and non-churners. This research applies several models where the most significant or best performance value is in the SVM model of 0.7996. The Neural Network model can be trained with better patterns to detect data and achieve high accuracy.

Keywords: Customer satisfaction, support vector machine, exploratoty data analysis, online store.

1. Introduction

Customers are the most critical asset in a business. Customers have many choices of services from several companies, so they can easily change services. The churn occurs when customers move from one store to another, tempted by more attractive offers, facilities, and prices [1]. Customer switching will cause the value flowing from the customer to decrease. If customer turnover continues, the company will gradually lose its competitive advantage. When the growth of new customers cannot meet the company's development needs, the company will avoid falling into a survival dilemma [2]. The company can retain many customers who move by implementing a timely sales marketing approach. One of the most direct and practical approaches to retaining current customers is for companies to be able to forecast potential churns in time and react quickly [3]. We cannot tolerate this churn phenomenon because it will decrease the company's turnover if we don't prevent and handle it.

However, most research on customer churn prediction concentrates on telecommunications, banking, retail, and other industries, leaving a gap in research on churn prediction for online stores. Online stores are trendy among the public, such as several marketplaces widely used in Indonesia, including Tokopedia, Bukalapak, Blibli, Zalora, Lazada, and other marketplaces [4]. Many things determine consumers to use online stores for shopping. Service quality determines consumer buying interest, leading them to purchase products in online stores.

Data mining is a new technology that is very useful in helping companies find important information from their data warehouse [5]. Data mining technology itself is at the core of knowledge discovery. This technology must sort out hidden and valuable knowledge from a large amount of data. In marketing decision-making, it

* **Author Correspondence**: Abu Tholib: Nurul Jadid University, PP Nurul Jadid road, Karanganyar, Paiton, Probolinggo, East Java 67291 – Indonesia. Email: ebuenje@gmail.com.

Copyright © 2024 The Authors, Published by Department of Informatics, UIN Saizu, Purwokerto – Indonesia
This work is licensed under Creative Commons Attribution 4.0 International.

is necessary to analyze enterprise data, find hidden rules, and further model them. Data mining has four processes: problem definition, data selection, model/algorithm sorting, and analysis results [6].

Support Vector Machine (SVM) is a relatively new technique compared to other methods. Still, it performs better in various application fields such as bioinformatics, handwriting recognition, text classification, etc. SVM is a machine learning model used to classify data into several classes with mathematically more mature concepts than Artificial Neural Networks (ANN) [7]. SVM is a guided learning method for classification. We use SVM for linear or nonlinear classification. We can find the best hyperplane by measuring the margin and identifying its maximum point. [8].

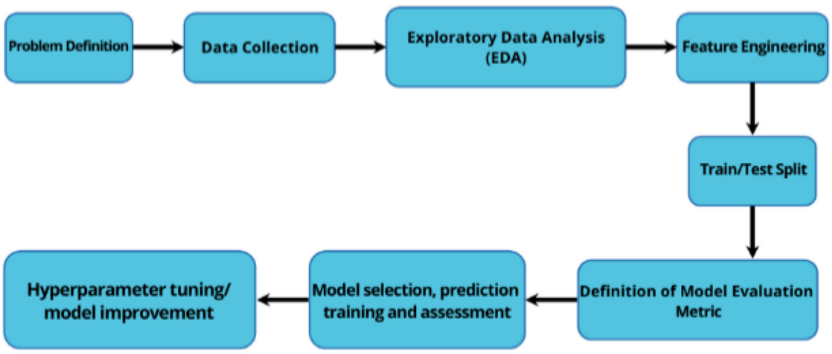
Previous research [9], [3], [2], [1] suggested that predicting customer churn is a challenging activity for every company because, most of the time, churn and non-churn customers have similar features. The churn prediction modeling system serves as a means to understand the exact behavior of customers and serves as a warning against the danger and timing of customer churn. The appropriateness of the strategy used is critical to achieving proactive retention intentions. After all, a company must anticipate a customer's intention to leave to make the right decisions about that customer. The study found that the longer the business relationship between the company and its customers lasts, the more profit the company will get from existing customers. Therefore, this churn model is critical to be raised in research.

Given the existing problems, we need techniques and problem-solving methods to identify why customers switch online stores. To identify these patterns, we use the SVM algorithm for problem-solving. This study confirms that the method can serve as an alternative model. This research aims to help predict the causes of customers switching to online stores. Companies can also improve customer service to minimize the number of moves.

2. Method

The research method was used to predict customer switching in online stores, as shown in Figure 1.

Figure 1.
Research steps



a. Problem Definition

Based on the introduction, the main challenge is predicting whether a customer will move. A machine learning model trains based on 80% of the sample data. We use the remaining 20% to apply the trained model and assess its predictive power for "churn" or "not churn." An additional question is, which features drive customers to switch? Such information can identify customer “pain points” and address them by providing things that can make customers stay.

b. Data Collection

The data used in this research is the Customer_churn-5 dataset. The data is available on Kaggle. This dataset consists of various attributes related to online store customers, which allows us to infer a consistent relationship between customer actions and churn.

c. Exploratory Data Analysis

Essentially, exploratory data analysis is an approach to seeing what the data can communicate to us, independent of formal modeling or hypothesis testing. EDA helps analyze a data set to summarize its statistical characteristics, focusing on four main aspects: measures of central tendency, measures of spread, shape of distribution, and presence of outliers. After data collection, several steps are taken to explore the data [10]. This step aims to understand the data structure, perform initial preprocessing, clean the data, identify patterns and inconsistencies in the data (e.g., skewness, outliers, missing values), and build and validate hypotheses.

d. Feature Engineering

Feature engineering techniques are usually applied after collecting and cleaning the input data. In the cleaning and feature selection steps, we remove redundant or unused features, resulting in a subset of the original features. It can be done before building the machine learning model or as part of the model building itself. Feature extraction reduces the dimensionality of the data set by creating new features. In principal component analysis, for example, the new features are linear combinations of the original features. Such dimension reduction techniques (which change the original variables) are often also applied or investigated during the feature engineering step. Many feature engineering techniques exist, and it is only sometimes clear which techniques fall under the definition of feature engineering and which do not.

e. Train-Test Split

We use the train-test split procedure to estimate the performance of machine learning algorithms when making predictions on data not used to train the model [11]. To perform the training and testing steps of the model, the data set is split into 80% training data and 20% testing data. The column "Churn" is defined as the class ("y"), and the other columns as features ("X").

f. Definition of Model Evaluation Metrics

Model evaluation is the process of using different evaluation metrics to understand the performance of a machine learning model [1], as well as its strengths and weaknesses. Model evaluation is essential for assessing model efficacy during the initial research phase and plays a role in model monitoring.

g. Model Selection, Training, Prediction, and Assessment

Our research used the SVM model. SVM is a linear classifier that classifies data by solving a quadratic optimization problem to determine the optimal separation plane between data sets. SVM can also be applied to nonlinear classification problems by kernel function transformation. SVM transforms the original problem into a more solvable dual problem through the Lagrange digger method, which is mathematically expressed in Formula 1 [4].

$$L(w,b)a) = \frac{1}{2}|w^2| + \sum_{i=1}^{m} a_i 1 - y_i\, w^T x_i + b \qquad (1)$$

SVM can simulate nonlinear decision boundaries well with kernel functions and control overfitting. The slower but accurate SVM model used here is in nonlinear form.

3. Results and Discussion

The implementation of this research has research results based on the stages of research that have been carried out. This research collects data obtained on Kaggle, which consists of 21 attributes and 7044 rows.

After data collection, several steps were taken to explore the data. This step aims to understand the data structure, perform initial preprocessing, clean the data, identify patterns and inconsistencies in the data (e.g., skewness, outliers, missing values), and build and validate hypotheses. Here is a table of some of the datasets we got on Kaggle, as shown in Table 1.

As a result of data preprocessing, the plot shows an imbalance of data classes between churners and non-churners. Resampling would be a suitable approach to address this. We retain the imbalance to simplify the case and select specific metrics for model evaluation. Churning customers have a significantly lower lifetime, with a median of about ten months, compared to the median of non-churning customers of about 38 months.

Table 1.
Dataset by Kaggle

Index	Customer ID	Gender	Partner	Dependents	Tenure	Phone Service
0	7590-VHVEG	Female	Yes	No	1	No
1	5575-GNVDE	Male	No	No	34	Yes
2	3668-QPYBK	Male	No	No	2	Yes
3	7795-CFOCW	Male	No	No	45	No
4	9237-HQITU	Female	No	No	2	Yes

Churning customers have a higher monthly bill with a median of about 80 USD and a much lower interquartile range than non-churning customers (median of about 65 USD)—total Cost results from lifetime and Monthly Cost, which is more individually insightful.

We apply the preprocessed data to an SVM model, which classifies the data by solving a quadratic optimization problem to establish the optimal separation plane between the data sets. SVM transforms the original problem into a more easily solved dual problem through the Lagrange multiplier method, which is mathematically expressed in Formula 1.

This study will test several models and measure their performance using several metrics. In the next step, we optimize the models by tuning their hyperparameters. The models used are KNN (K Nearest Neighbors), Logistic Regressions, Random Forest, and SVM.

We use cross-validation during hyperparameter tuning with Grid Search and Randomized Search to overcome the potential bias from the specific data splitting in the train-test-split section. Cross-validation divides the training data into a particular number of folds. We keep one-fold as the "training-dev" set for each iteration and use the other as the training set. The cross-validation result is the k-value for all metrics on the k-fold CV. For KNN, GridSearchCV determines the optimal number of neighbors (k) that results in the best model performance. For Logistic Regression, GridSearchCV is used to determine the best model while applying different values of L1 or L2 regularization to change the impact of meaningless features to zero (L1) or to simplify the model by relativizing strong patterns picked up during training (L2). For the Random Forest model, RandomizedSearchCV was used to optimize several hyperparameters, including n_estimator, max_features, max_depth, criterion, and bootstrap. For SVM, GridSearchCV determines the C value for the optimal margin around the support vector by comparing accuracy values on several models, as shown in Table 2.

Table 2.
Comparison of the Accuracy Level of Each Model

Model	Accuracy Value
KNN	0.7633
Logistic Regression	0.7931
Random Forest	0.7782
SVM	0.7996

In this study, we also added a SVM model. Although the data set is relatively small and networks generally require training data to develop meaningful prediction capabilities, a simple neural network is used for quick comparison with other approaches, as shown in Figure 2.

Figure 2.
Training and Testing Accuracy Graph

Figure 3.
Evaluation of Matric Value

```
Confusion Matrix:
[[918 115]
 [167 207]]
---------------------------------------
Accuracy: 0.799574
ROC AUC: 0.826234
---------------------------------------
Precision: 0.642857
Recall: 0.553476
F1 score: 0.594828
```

As shown in Figure 3, the best accuracy on the test set was achieved by the SVM with a value of 0.7996, given the high data imbalance towards non-churners. It makes sense to compare the F1 values to get the model with the best value in terms of precision and recall. The model is a SVM with an F1 value of 0.5948. Further optimization efforts should be made to achieve higher scores and thus increase the predictive power for larger business values.

Online stores typically include more data in the analysis, such as extended customer and transaction data from CRM systems and operational data from network services. They also typically have much larger churn/non-churn events than a CA. 7000 in this example case. We can adequately train a neural network model to detect more complex patterns in the data and achieve higher accuracy. High accuracy is necessary to identify promising customer cases where churn can be avoided because the return of the protected customer needs to outweigh the cost of the associated retention campaign.

4. Conclusion

This research applies several models to predict customer switching for the best results. The KNN model gets an accuracy value of 0.7633, Logistic Regression of 0.7931, Random Forest of 0.7782, support vector machine of 0.7938. So, the SVM model gets the best value compared to other models with the best performance value with an F1 score.

SVM models can be adequately trained to detect more complex patterns in the data and achieve higher accuracy. High accuracy is necessary to identify promising customer cases where churn can be avoided because, ultimately, the return of a protected customer needs to outweigh the cost of the associated retention campaign.

5. Acknowledgment

We want to thank Nurul Jadid University for the facilities and facilities they provide, which have supported the smooth implementation of experiments and data analysis. We would also like to thank the lecturers who helped us prepare this scientific article.

6. Reference

[1] V. R. R. Raj and R. A. Azad . V, “Customer Churn Prediction in Telecommunication Industry Having Data Certainty,” *Int. J. Sci. Res. Sci. Eng. Technol.*, pp. 113–122, Jul. 2020, doi: 10.32628/IJSRSET207427.

[2] M. Zhao, Q. Zeng, M. Chang, Q. Tong, and J. Su, “A Prediction Model of Customer Churn Considering Customer Value: An Empirical Research of Telecom Industry in China,” *Discrete. Dyn. Nat. Soc.*, vol. 2021, pp. 1–12, Aug. 2021, doi: 10.1155/2021/7160527.

[3] I. M. M. Mitkees, S. M. Badr, and A. I. B. ElSeddawy, “Customer Churn Prediction Model Using Data Mining Techniques,” in *2017 13th International Computer Engineering Conference (ICENCO)*, Cairo: IEEE, Dec. 2017, pp. 262–268. doi: 10.1109/ICENCO.2017.8289798.

[4] 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.

[5] I. Ernawati, “Data Mining Sebagai Salah Satu Solusi Strategi Bisnis,” *Inform. J. Ilmu Komput.*, vol. 14, no. 1, p. 9, Apr. 2018, doi: 10.52958/iftk.v14i1.367.

[6] Y. Sun and X. Tan, “Customer Relationship Management Based on SPRINT Classification Algorithm under Data Mining Technology,” *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/6170335.

[7] B. A. Nugroho, A. K. A. Pradana, E. Nurfarida, and and, “Prediksi Waktu Kedatangan Pelanggan Servis Kendaraan Bermotor Berdasarkan Data Historis menggunakan Support Vector Machine,” *JEPIN (Jurnal Edukasi dan Penelit. Inf.*, vol. 7, no. 1,p, p. 25, 2021.

[8] F. Handayani, “Komparasi Support Vector Machine, Logistic Regression dan Artificial Neural Network dalam Prediksi Penyakit Jantung,” *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, p. 329, Dec. 2021, doi: 10.26418/jp.v7i3.48053.

[9] S. B, S. M, and N. D, “Customer Churn Prediction,” *IARJSET*, vol. 8, no. 6, pp. 527–531, Jun. 2021, doi: 10.17148/IARJSET.2021.8692.

[10] K. Sahoo*, A. K. Samal, J. Pramanik, and S. K. Pani, “Exploratory Data Analysis using Python,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 4727–4735, Oct. 2019, doi: 10.35940/ijitee.L3591.1081219.

[11] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. Vanden Broucke, “Special issue on feature engineering editorial,” *Mach. Learn.*, vol. 113, no. 7, pp. 3917–3928, Jul. 2024, doi: 10.1007/s10994-021-06042-2.

This page is intentionally left blank