



Using Machine Learning Algorithms to Analyze Customer Churn in the Software as a Service (SaaS) Industry

*¹Levent ÇALLI, ²Sena KASIM

¹Department of Management and Organization, Sakarya University, Sakarya, Turkey, lcalli@sakarya.edu.tr 

²Department of Information Systems Engineering, Sakarya University, Sakarya, Turkey, senakasim@gmail.com 

Abstract

Companies must retain their customers and maintain long-term relationships in industries with intense competition. Customer churn analysis is defined in the literature as identifying customers who may leave a company to take appropriate marketing precautions. While customer churn research is prevalent in B2C (Business to Customer) business models such as the telecoms and retail sectors, customer churn analysis in B2B (business to business) models is a relatively emerging topic. In this regard, the study carried out a customer churn analysis by considering an ERP (enterprise resource planning) company with a software as a service (SaaS) business model. Different machine learning algorithms analyzed ten features determined by selection methods and expert opinions. According to the analysis results, the random forest algorithm gave the best result. Additionally, it has been observed that the number of products and customer features has a relatively higher weight for the prediction of churner.

Keywords: Customer Churn, SaaS, Machine Learning, Random Forest, Data Mining

1. INTRODUCTION

In an era where the reduction of costs is an essential factor coupled with intense competitive pressure, organizations must fully maximize their existing customer; therefore, customer retention efforts are implemented to detect decreasing loyalty of the customers with churn analysis [1]. Dissatisfaction, high cost, poor quality, lack of features, privacy concerns, or many different features can be caused the loss of customers. Hence, identifying these features that change according to the industries that decrease loyalty of customers and making reasonable efforts by companies will contribute to a positive change in the situations of customers who are likely to churn [2]. Consequently, loyal customers make more purchases, pay a premium price, and acquire new customers through favorable word-of-mouth, which positively impacts the company's long-term reputation [3]. Any company that wants to survive in business cannot simply ignore the churner and loyalty concept.

Churn analysis has been carried out in different industries in the academic literature. The most intense studies focus on B2C business models such as telecommunication [2], [4]–[10], financial services [11]–[14], and retail [15]–[17] industries. However, few studies have been conducted considering the B2B business model. In B2B markets, fewer

customers commonly make larger and more frequent purchases, for instance, compared with the telecommunication or supermarket industries. Hence, customers are more valuable for companies operating the B2B model, and customer retention is central to developing long relationships [18].

In this regard, this study aims to fill the literature gap with a customer churn approach considering the SaaS (software as a service) industry. This research analyzed customer churn using various machine learning algorithms in a cloud ERP company. Furthermore, feature selection techniques were used, and more effective ones for customer churn analysis were identified. The research findings are expected to have academic and practical benefits, particularly in the SaaS field, where few studies exist.

2. LITERATURE REVIEW

2.1. Churn Analysis

Today, our ability to store and analyze all types of data as a result of the information society due to the rapid development of information and communication technologies (ICT) in recent years has resulted in the acquisition of valuable knowledge through data mining

* Corresponding Author

methods [19]. Customer churn analyses using the data collected from the customer are often used today within the framework of the loyalty concept, which is vital for all businesses to achieve long-term relationships with their customers. In this way, companies can identify customers with the potential to leave them using various data mining and machine learning methods to effort the necessary marketing activities for this group called churners.

In the literature, companies using the B2C business model have commonly been analyzed using the customer churn

approach. Telecom is the most widely researched industry in this field. For example; Huang et al. [4] indicate that different features and machine learning methods can effectively analyze customer churn. Ahn et al. [7] show that customer transaction and billing data are important factors for customer churn in mobile telecom services. Verbeke et al. [6] present that a small number of datasets can predict churning with high accuracy and Bhattacharyya and Dash [10] discuss the customer churn analysis in the telecom industry from a bibliometric perspective.

Table 1. Academic literature review on churn analysis in SaaS

Study	Methods	Feature Selection	Features and Data Size	Sector	Findings
[20]	<ul style="list-style-type: none"> • Logistic Regression • Random Forest • XGBoost 	No	21 Features 8,256 Observations	No Information	Logistic Regression and XGboost performed relatively well with %72 and %75 AUC scores, respectively. The number of Type-A User Login is the feature with the highest weight.
[21]	<ul style="list-style-type: none"> • Long short-term memory (LSTM) • Convolutional Neural Network (CNN) • Support Vector Machine • Random Forest 	No	5 Feature Categories No data size information	Advertising	Random Forest and Support Vector Machine performed best with 83% and %81 accuracies, respectively. The LSTM and CNN, as the deep learning methods, performed poorly due to a lack of data. As platform usage data, the customer's minutes spent on the platform and the number of active users carries the most weight.
[22]	<ul style="list-style-type: none"> • Logistic Regression • Support Vector Machine • Decision Tree • Random Forest 	Chi Square Anova	23 Features 1788 Observations	Inventory Management	Random Forest performed best with 92% accuracy The number of transactions is the feature with the highest weight.
[23]	<ul style="list-style-type: none"> • Logistic Regression • Random Forest 	No	43 Feature under 4 Categories 8869 Observations Random oversampling and undersampling methods were used.	Cloud-based business phone system and call center	Successful results with little data could not be obtained for both algorithms considering precision and recall ratios. The features with the highest weight are the number of users, number of integrations and call quality.

The customer churn study conducted for the banking industry shows that customers who use more banking services have become more loyal. Customers who use less than three services are the group that needs attention [13]. In the study conducted by Keramati et al. [12] in the banking industry, variables such as the number of mobile, internet, and telephone bank transactions and demographic variables such as age, gender, and educational differences are influential on customer churn prediction.

While customer churn analysis-related studies conducted in the field of B2C are common in the literature, there are few studies on companies using the B2B business model, where the SaaS companies are one of them. The following section discusses the concept of SaaS and research done in this area.

2.2. Software-as-a-Service (SaaS)

SaaS is a business model that offers cloud-based services to clients as an alternative to standalone software that requires installation, maintenance, IT infrastructure, and support services (backup, upgrade, security), especially for B2B [24]. According to Fortune Business Insights (2022)'s report [25] and Jones [27], the global SaaS marketplace is expected to grow from \$130 billion in 2021 to \$716 billion in 2028, with Microsoft, Salesforce, Adobe, SAP, and Oracle accounting for 51% of the market. Since the market's growth

potential will bring more competition, it is vital to determine the factors clients consider in their SaaS preferences to retain and gain new customers. Allen [28] states that a SaaS firm's 3% and 8% monthly customer churn is average. In this respect, monitoring customer churn rates through data mining approaches and intervention to prevent loss are essential for this highly competitive market.

In the academic literature, churn analysis studies tend to focus on the telecommunications, financial services, and retail sectors as B2C business models, while there are just a few studies in the SaaS industry as a B2B business model. These studies are shown in Table 1.

Ge et al. [20] performed churn analysis using logistic regression, random forests, and XGBoost algorithms in their study conducted with the data of a SaaS company, considering 8256 observations and 21 features. They observed that Logistic regression and XGBoost algorithms made relatively good predictions for churning. Additionally, it has been discovered that online usage behaviors, such as login and project numbers, strongly predict customer churn.

In another study conducted in the scope of SaaS, Rautio [21] performed churn analysis with different machine learning methods, considering a company operating in the advertising industry. The research's features mainly focused on client

business data, such as spending, platform usage, previous customer service interactions, and service-related customer feedback. In the study, where the support vector machine algorithm gave the best results, it was determined that platform usage metrics feature relatively more weight in prediction. The study also found that deep learning methods did not perform well in predicting churner.

Amornvetchayakul and Phumchusri [22] used four machine learning algorithms in their study with 1778 samples for churn analysis, considering a SaaS company that provides inventory management services (inventory levels, purchase orders, delivery, etc.) for SMEs (Small and Medium Enterprises). The number of transactions in the current month and the number of transactions in the previous month features were discovered as having considerably higher weights than other features in their research, which considered 23 features. The prediction of the random forest algorithm has stated that it has higher accuracy than the decision tree, logistic regression, and support vector machine algorithms.

Lastly, a study focused on churn analysis of a firm in the cloud-based business phone system and call center industry

revealed that neither the random forest approach nor logistic regression could achieve sufficient accuracy rates due to strongly overfitting [23]. The number of users, number of integrations, and call quality features were shown to be more critical in predicting churners with the random forest algorithm when random oversampling and undersampling approaches were utilized.

3. METHOD

This research focused on a software company based in Germany and Turkey that specialized mainly in the ERP (enterprise resource planning) industry. A total of 1951 observations were analyzed. Initially, the weight of sixteen features was evaluated, and then the prediction phase that considered churn and non-churn customers was performed using various machine learning methods. The feature selection stage used the Chi-square, information gain, gain ratio, and Gini index methods. Decision Tree, Logistic Regression, Naive Bayes, K-NN, Random Forrest, and Neural Networks algorithms were utilized as classification methods. This process is shown in Figure 1.

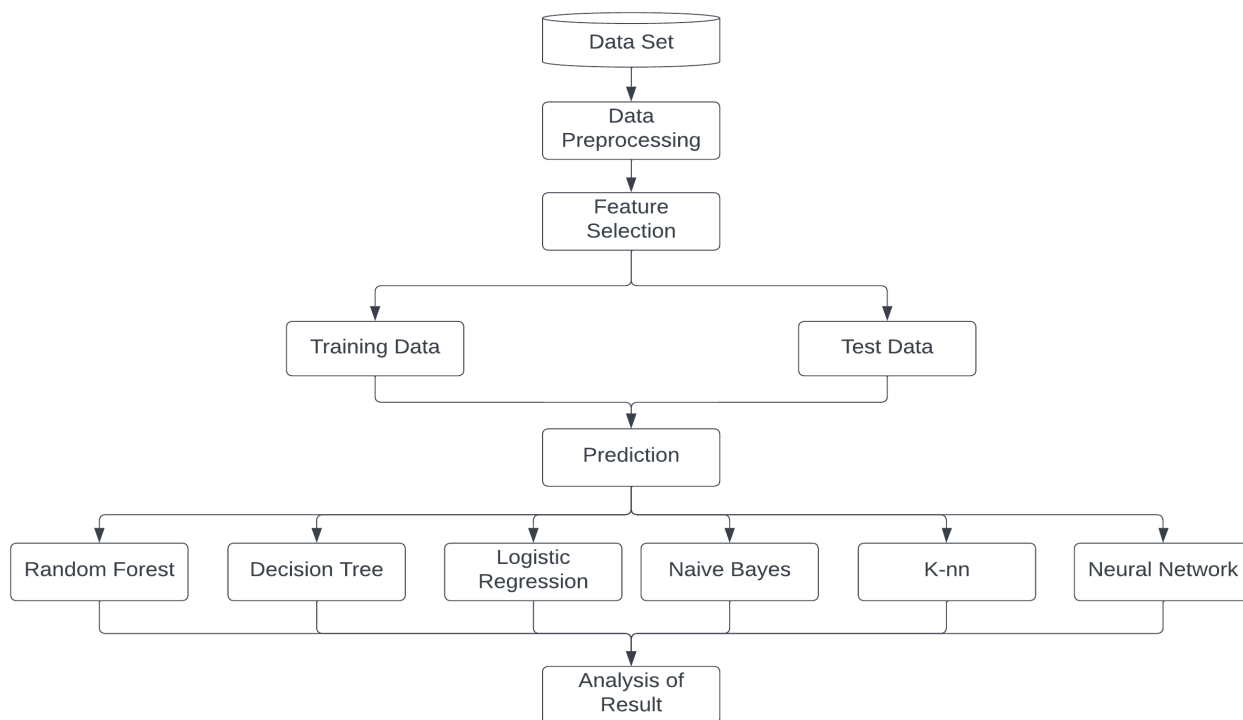


Figure 1. Research process

3.1. Feature Selection

In data mining, feature selection is the preferred technique to reduce dataset size to achieve more efficient analysis and adapt the dataset better to match the preferred analysis method [29]. This study used the chi-square, information gain, gain ratio, and Gini index methods to determine the features to be considered in the churn analysis.

3.1.1. Chi-Square

The Chi-square test, a non-parametric method, is used to examine whether there is a relationship between two

categorical variables [30]. The formula for the Chi-square is as follows.

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

In formula (1), c represents the degrees of freedom, O the observed values, and E the expected value [31].

3.1.2. Information Gain and Gain Ratio

Information gain is used to identify the best features that provide the most information about a class and uses the idea

of entropy which is defined as a measure of purity or the degree of uncertainty of a random variable [32]. The information gain calculates the entropy difference before and after the division and determines the purity of the in-class elements. The entropy is calculated using the following formula (2) [33];

$$H(S) = - \sum_{i=1}^N p_i \log_2 p_i \quad (2)$$

S: Set of all examples in the dataset
N: Number of distinct class values
p_i: Event probability

The information gain is calculated with the formula (3) shown below [33];

$$\begin{aligned} \text{Information Gain}(A, S) &= H(S) - \sum_{j=1}^v \frac{|S_j|}{|S|} \cdot H(S_j) \\ &= H(S) - H(A, S) \end{aligned} \quad (3)$$

H(S): Entropy of the whole dataset *S*
|S_j|: Number of the instance with *j* value of an attribute *A*
|S|: Total number of instances in dataset *S*
v: Set of distinct values of an attribute *A*
H(S_j): Entropy of subset of instances for attribute *A*
H(A, S): Entropy of an attribute *A*

The gain ratio method is estimated by dividing the information gain value by the entropy value to get accurate results due to the asymmetrical nature of the information gain results [34].

Table 2. Feature weight ranking by selection methods

Ranking	Chi-square	Gini Index	Information Gain	Gain Ratio
1	Number of Invoices	The number of customers	The number of customers	Number of Cash Register Connections
2	Number of Offers	Number of Invoices	Number of Invoices	The number of customers
3	Number of Support	Number of products	Number of Offers	Custom Report Usage
4	The number of customers	Number of Offers	Number of Support	Mail Connection
5	Number of products	Number of Support	Number of products	Number of Offers
6	Number of Cash Register Connections	Number of Users	Number of Users	Number of Users
7	Cargo Usage	Cargo Usage	Number of Cash Register Connections	Number of Invoices
8	Number of Payment Documents	Number of Cash Register Connections	Cargo Usage	Number of Support
9	Custom Report Usage	Custom Report Usage	Custom Report Usage	Cargo Usage
10	Mail Connection	Number of Payment Documents	Number of Payment Documents	Number of Payment Documents
11	Number of Cash Register Receipts	Number of Orders	Mail Connection	Number of products
12	Number of Users	Number of Cash Register Receipts	Number of Orders	Number of Cash Register Receipts
13	Number of Orders	Mail Connection	Number of Cash Register Receipts	Number of Production Orders
14	Customer Group	Customer Group	Customer Group	Number of Orders
15	Number of Production Orders	Number of Production Orders	Number of Production Orders	Customer Group
16	Number of Marketplaces	Number of Marketplaces	Number of Marketplaces	Number of Marketplaces

3.1.3. Gini Index

As an impurity splitting method, Gini Index is appropriate for binary and continuous numeric values for calculating feature weight. Assume that *S* is the collection of *s* samples with *m* different classes ($C_i, i=1, \dots, m$). *S* can be divided into *m* subsets based on class differences ($S_i, i=1, \dots, m$). If S_i is the sample set for class C_i , and S_i is the number of samples in S_i , then the Gini index of *S* is calculated with the following formula (4) [35];

$$\text{Gini Index}(S) = 1 - \sum_{i=1}^m P_i^2 \quad (4)$$

In this formula, P_i refers to the probability that each given sample belongs to C_i and is measured with s_i/s . When Gini Index has a minimum value of 0, it indicates that all members of the set fall under the same class, indicating that it can gather the most valuable information [35]. IBM SPSS and Orange Data Mining software were utilized for feature selection algorithms [36], [37].

Table 2 shows the ranking of features according to the four feature selection methods. Ten key features were selected for churn analysis within the scope of analysis results and expert opinion.

3.2. Predictive Analysis Algorithms

Churn analysis was performed with Decision Tree, Logistic Regression, Naive Bayes, K-NN, Random Forest, and Neural Networks algorithms which are common approaches in the literature for this study. Each algorithm shortly explains in this section.

3.2.1. Decision Tree

The decision tree is one of the most popular data categorization techniques in literature, which is based on the information gain theory explained in the previous section. Nodes and branches create the decision tree structure. Nodes represent tests on certain features, while branches indicate test results.

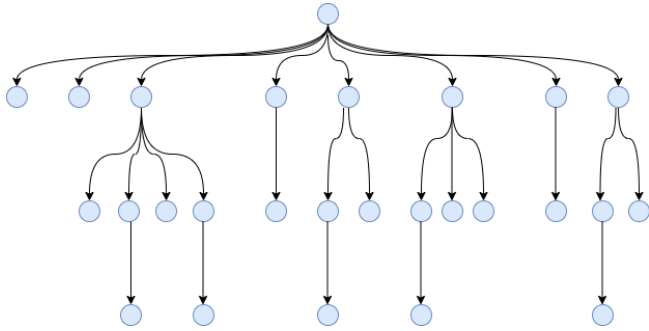


Figure 2. Decision Tree model

Although the decision tree technique has several alternative approaches, the most well-known C5 method splits the sample using the feature with the maximum information gain. It then repeats this process until the subset can no longer be divided [17]. An example decision tree model is shown in Figure 2.

3.2.2. Logistic Regression

Logistic regression is a commonly used statistical model for predicting event probability. In the Logistic regression model seen in detail in formula 5, the dependent variable y is a binary in the formula used to determine whether an event occurred.

$$\text{prob}(y = 1) = \frac{e^{\beta_0 + \sum_{k=1}^K \beta_k x_k}}{1 + e^{\beta_0 + \sum_{k=1}^K \beta_k x_k}} \quad (5)$$

The independent inputs are x_1, x_2, \dots, x_k . The maximum likelihood approach can estimate $\beta_1, \beta_2, \dots, \beta_k$ as the regression coefficients based on the available training data [4].

3.2.3. Naïve Bayes

The Naive Bayes algorithm is a well-known classification method used in the machine learning literature. It has attracted much interest due to its ease of use and good performance [38]. The formula (6) is shown below.

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)} \quad (6)$$

Bayes theorem estimates the posterior probability of class given predictor, $P(c|x)$, from $P(c)$ (the prior probability of a class), $P(x)$ (the prior probability of predictor), and $P(x|c)$ (represents the likelihood, which is the probability of predictor class given) [39].

3.2.4. K-NN (The K-Nearest Neighbors Algorithm)

The K-Nearest Neighbors algorithm is an easy-to-implement, simple with few hyperparameters required, supervised learning algorithm that produces classifications or predictions for clustering of a single data point using proximity techniques such as Euclid, Manhattan, Minkowski, and Hamming [40].

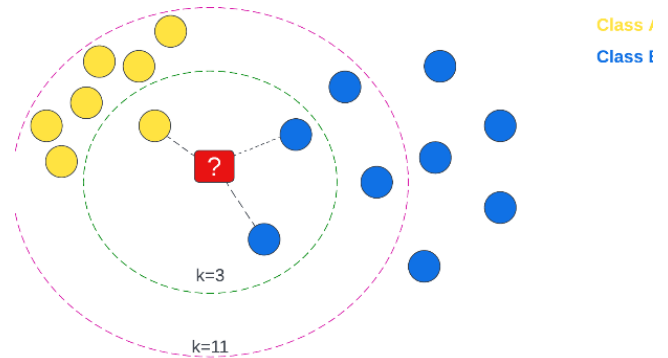


Figure 3. The K-Nearest Neighbors algorithm

K-NN uses most k nearest neighbors for a new data point whose class is being looked for to assign [41]. This situation is seen in Figure 3. For instance, the red shape containing an unclassified data point may belong to either class A or B according to the alternative k values.

3.2.5. Random Forest

The random forest algorithm is applied to a wide range of prediction problems as a well-known approach in the literature due to its capacity to handle small sample sizes and as well as high-dimensional feature spaces, with a few parameters for tuning [42]. The random forest comprises many independent decision trees that act as an ensemble method, and each tree in the random forest generates a class prediction. The model's prediction is based on the class with the most votes, as seen in Figure 4 [43].

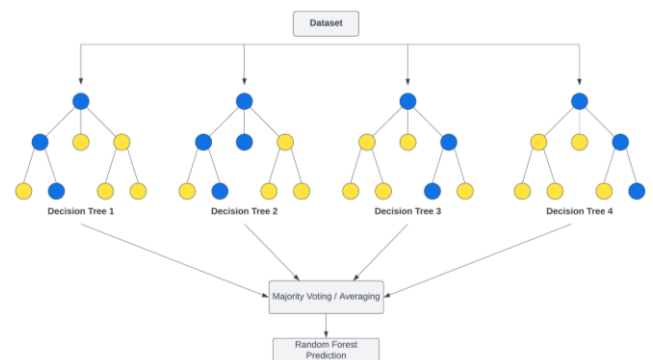


Figure 4. Random Forest

3.2.6. Neural Networks

Neural networks are a method of explaining cognitive, decision-making, and other intelligent control behaviors by using the way the human brain operates as a kind of data processing and analysis [17]. A classic neural network consists of three layers: an input layer, a hidden layer, and an output layer, all connected by neurons, as seen in Figure 5.

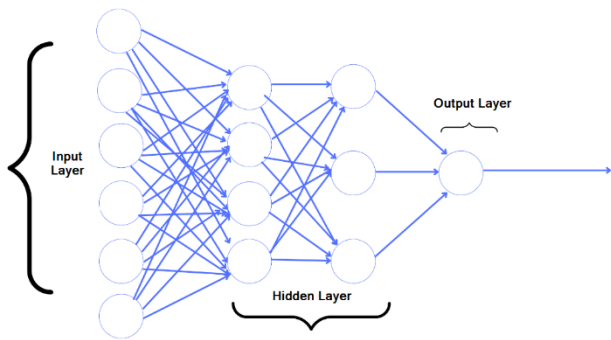


Figure 5. Neural Networks

A key advantage of utilizing neural networks for data modeling is that they can fit complicated nonlinear models without needing to be specified in advance, unlike other nonlinear estimation methods [44].

4. RESULTS

As a result of the feature selection analysis stage and considering the opinions of industry experts, it was determined that ten features would be appropriate for churn analysis. Table 3 shows these features and their descriptions.

Table 3. Selected features

Features	Description
Number of products	The number of products registered in the customer’s account.
Number of Customers	The total number of customers and suppliers listed on the customer’s account.
Number of Offers	The number of offers that the customer creates in the account.
Number of Orders	The number of orders created by the customer.
Number of Invoices	The number of invoices created by the customer.
Cargo Usage	The number of cargo companies the customer has used in the account.
Number of Users	The number of users the customer has in the ERP software that can access the system at the same time.
Custom Report Usage	The reports are specially made for the customer, excluding the standard reports in the ERP software.
Number of Cash Register Receipts	The number of cash register receipts created by the customer in the ERP software.
Email Connection	Email connection status used for the proposal side of the customer’s ERP software.

The churn status of customers was considered in this study based on their active usage of the software. 836 customers in the data set are coded as churn (0), and 1115 are active

customers (1). In this regard, it is thought that the dependent variable has a balanced distribution.

Correlation analysis, an essential part of descriptive statistics, was done in Python to reveal the relationships between variables. Figure 6 shows the correlation of features with each other and the customer churn as the dependent variable. A strong positive correlation is shown by dark colors, whereas light tones indicate a weak positive correlation.

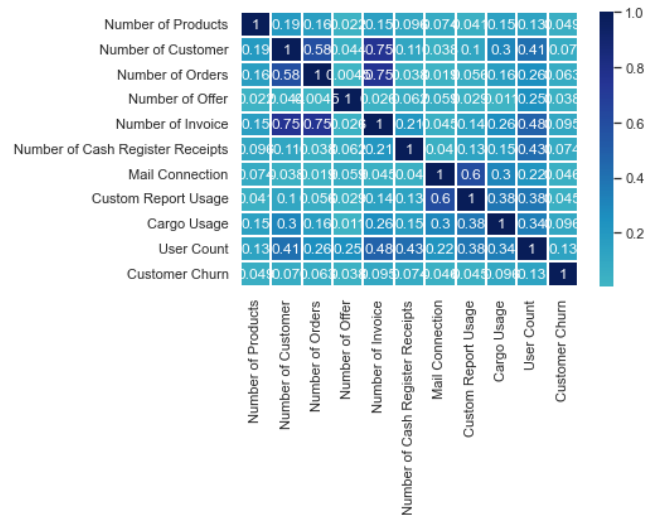


Figure 6. Correlation heatmap

The normalization process was conducted in the next step by assigning values between 0 and 1 due to the features’ value variations. This method allows using a single scale while keeping the distinctions in the value ranges and avoiding information loss of each feature.

Table 4. Number of training and test data

	Training Dataset	Test Dataset
%75-%25	1463	488
%70-%30	1365	586

In the next stage, the dataset was divided into test and training data. Then, the model was tested with different algorithms. This study used two different ratios of training data, 70% and 75%. The number of customers that create the training and test data sets is shown in Table 4.

Table 5. Churn Analysis results

Predictive Algorithms	%75-%25	%70-%30
	Accuracy	Accuracy
Decision Tree	74.59%	73.54%
Logistic Regression	57.58%	57.84%
Naive Bayes	47.95%	48.12%
K-NN	65.36%	74.23%
Random Forest	78.27%	77.47%
Neural Network	57.99%	58.19%

As seen in Table 5, the churn analysis results with the training and test data in both ratios show that the random forest algorithm gives the relatively best results with higher accuracy rates. The results also show that the decision tree

algorithm produces better results than other algorithms. Additionally, the K-NN method provides much better accuracy when the number of training data sets is decreased. The weights of the features used in the predictive algorithms for the random forest and decision tree are shown in Table 6.

Table 6. Features weight

Decision Tree		Random Forest	
Feature	Weight	Feature	Weight
Number of Customer	45,0%	Number of Customer	42,4%
Number of Products	29,6%	Number of Product	33,3%
Number of Invoice	15,4%	Number of Invoice	9,4%
Number of Orders	4,0%	Number of Orders	5,4%
Number of Users	2,2%	Number of Offer	3,1%
Cargo Usage	1,7%	Cargo Usage	2,8%
Number of Offer	1,3%	Number of Users	2,8%
Custom Report Usage	0,5%	Number of Cash Register Receipts	0,3%
Mail connection	0,2%	Mail Connection	0,3%
Number of Cash Register Receipts	0,0001%	Custom Report Usage	0,3%

In this respect, it is seen that the number of customers and products has a high weight in both algorithms predicting the customer's churn status. Mail connection, custom report usage, number of cash register receipts, cargo usage, number of the offer, number of users, and number of orders have low weight, while the number of invoice feature has a relatively moderate weight in both algorithms.

5. CONCLUSION

Customer churn is a severe challenge to any business, and one way to deal with it is to predict which customers are most likely to leave the company and then target those segments with marketing efforts to encourage them to be loyal [45]. Customer loyalty is valuable since acquiring new customers is more costly than keeping existing customers, and a loyal customer serves as an honorary lawyer for the company, becoming a positive reference for potential customers and, as a result, generating more profit [19]. While academic studies on customer churn analysis are popular in a few industries, especially in B2C fields such as telecommunications, grocery retail, or bank, research on software as a service (SaaS) as a B2B industry is very limited. Thus, the study aimed to fill a gap in the academic literature by focusing on a cloud ERP business.

According to the research findings, it is seen that the number of customers and the number of products are the most important features in customer churn analysis. The number of invoices and orders features also shows a relatively higher weight than others. Thus, research findings reveal that features of the fundamental business process have a meaningful relationship with customer churn and are factors that need more attention for managers. In this sense, the results show some parallelism with Amornvetchayakul and

Phumchusri [22] and Sergue [23]. For example, Amornvetchayakul and Phumchusri [22] indicate that the importance of the number of transactions in the inventory management sector is relatively higher than in others. Additionally, Sergue [23] highlights that the number of users in the Cloud-based business phone system and call center industry is relatively higher weight than other features in predicting churner. In terms of prediction accuracy, the random forest algorithm achieves the best result in this study, which is similar to the results observed by Rautio [21] and Amornvetchayakul and Phumchusri [22].

As a result, the churn analysis processes of B2B business models, which have less data for many businesses compared to B2C business models, undoubtedly involve many difficulties. In this study, the customer churn analysis process was carried out for a cloud ERP company and contributed to the academic literature and the practical field. Although it is an important gap that there are very few studies in this literature, it is thought that new studies will emerge in the light of the findings of this study. Using different features and testing different predictive algorithms will enrich the academic literature and be guiding for managers.

Author Contributions: Concept – L.C., S.K.; Data Collection &/or Processing – S.K; Literature Search – L.C., S.K.; Writing – L.C.

Conflict of Interest: This study was produced from the master's thesis entitled "Customer Churn Analysis with Data Mining Methods: Software Industry," conducted at Sakarya University, Institute of Natural Sciences, with the supervisor Levent ÇALLI.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

- [1] N. Gladly, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," *Eur. J. Oper. Res.*, vol. 197, no. 1, pp. 402–411, 2009, doi: 10.1016/j.ejor.2008.06.027.
- [2] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *J. Bus. Res.*, vol. 94, no. February 2018, pp. 290–301, 2019, doi: 10.1016/j.jbusres.2018.03.003.
- [3] J. Ganesh, M. J. Arnold, and K. E. Reynolds, "Understanding the customer base of service providers: An examination of the differences between switchers and stayers," *J. Mark.*, vol. 64, no. 3, pp. 65–87, 2000, doi: 10.1509/jmkg.64.3.65.18028.
- [4] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1414–1425, 2012, doi: 10.1016/j.eswa.2011.08.024.
- [5] K. Kim, C. H. Jun, and J. Lee, "Improved churn prediction in telecommunication industry by analyzing a large network," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6575–6584, 2014, doi: 10.1016/j.eswa.2014.05.014.

- [6] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, 2012, doi: 10.1016/j.ejor.2011.09.031.
- [7] J. H. Ahn, S. P. Han, and Y. S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," *Telecomm. Policy*, vol. 30, no. 10–11, pp. 552–568, 2006, doi: 10.1016/j.telpol.2006.09.006.
- [8] C. F. Tsai and Y. H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12547–12553, 2009, doi: 10.1016/j.eswa.2009.05.032.
- [9] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzivasvas, "A comparison of machine learning techniques for customer churn prediction," *Simul. Model. Pract. Theory*, vol. 55, pp. 1–9, 2015, doi: 10.1016/j.simpat.2015.03.003.
- [10] J. Bhattacharyya and M. K. Dash, "What Do We Know About Customer Churn Behaviour in the Telecommunication Industry? A Bibliometric Analysis of Research Trends, 1985–2019," *FIIB Bus. Rev.*, 2021, doi: 10.1177/231971452111062687.
- [11] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 5445–5449, 2009, doi: 10.1016/j.eswa.2008.06.121.
- [12] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financ. Innov.*, vol. 2, no. 1, 2016, doi: 10.1186/s40854-016-0029-6.
- [13] A. Bilal Zoric, "Predicting Customer Churn in Banking Industry using Neural Networks," *Interdiscip. Descr. Complex Syst.*, vol. 14, no. 2, pp. 116–124, 2016, doi: 10.7906/indecs.14.2.1.
- [14] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 4626–4636, 2009, doi: 10.1016/j.eswa.2008.05.027.
- [15] M. A. de la Llave Montiel and F. López, "Spatial models for online retail churn: Evidence from an online grocery delivery service in Madrid," *Pap. Reg. Sci.*, vol. 99, no. 6, pp. 1643–1665, 2020, doi: 10.1111/pirs.12552.
- [16] W. Buckinx and D. Van Den Poel, "Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," *Eur. J. Oper. Res.*, vol. 164, no. 1, pp. 252–268, 2005, doi: 10.1016/j.ejor.2003.12.010.
- [17] X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," *2020 IEEE 5th Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2020*, pp. 129–132, 2020, doi: 10.1109/ICCCBDA49378.2020.9095611.
- [18] B. Janssens, M. Bogaert, A. Bagué, and D. Van den Poel, "B2Boost: instance-dependent profit-driven modelling of B2B churn," *Ann. Oper. Res.*, 2022, doi: 10.1007/s10479-022-04631-5.
- [19] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2354–2364, 2011, doi: 10.1016/j.eswa.2010.08.023.
- [20] Y. Ge, S. He, J. Xiong, and D. E. Brown, "Customer churn analysis for a software-as-a-service company," in *2017 Systems and Information Engineering Design Symposium, SIEDS 2017*, 2017, pp. 106–111, doi: 10.1109/SIEDS.2017.7937698.
- [21] A. Rautio, "Churn rediction in SaaS using Machine Learning," 2019.
- [22] P. Amornvetchayakul and N. Phumchusri, "Customer Churn Prediction for a Software-as-a-Service Inventory Management Software Company: A Case Study in Thailand," in *2020 IEEE 7th International Conference on Industrial Engineering and Applications, ICIEA 2020*, 2020, pp. 514–518, doi: 10.1109/ICIEA49774.2020.9102099.
- [23] M. Sergue, "Customer Churn Analysis and Prediction using Machine Learning for a B2B SaaS company," 2020, [Online]. Available: www.kth.se/sci.
- [24] D. Ma, "The Business Model of 'Software-As-A-Service,'" in *IEEE International Conference on Services Computing (SCC 2007)*, 2007, no. July, pp. 701–702, doi: 10.1109/SCC.2007.118.
- [25] Fortunebusinessinsights, "The software as a service market Size," 2022. <https://www.fortunebusinessinsights.com/software-as-a-service-saas-market-102222>.
- [26] E. Jones, "Cloud Market Share – a Look at the Cloud Ecosystem in 2022," *KINSTA BLOG*, 2022. <https://kinsta.com/blog/cloud-market-share/#:~:text=The SaaS market is dominated,impressive annual growth of 34%25>.
- [27] E. Jones, "Cloud Market Share – a Look at the Cloud Ecosystem in 2022," *KINSTA BLOG*, 2022. .
- [28] K. Allen, "Churn Rate vs Retention Rate: How to Calculate These SaaS KPIs," *woopra.com*, 2022. <https://www.woopra.com/blog/churn-rate-vs-retention-rate>.
- [29] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, no. May, pp. 1200–1205, 2015, doi: 10.1109/MIPRO.2015.7160458.
- [30] A. Field, *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications Ltd, 2018.
- [31] sampath kumar Gajawada, "Chi-Square Test for Feature Selection in Machine learning,"

- <https://towardsdatascience.com/>, 2019.
<https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223>.
- [32] N. Tyagi, "What is Information Gain and Gini Index in Decision Trees?," <https://www.analyticssteps.com/>, 2021. <https://www.analyticssteps.com/blogs/what-gini-index-and-information-gain-decision-trees>.
- [33] U. Krčadinac, "Classification – Decision Trees," 2015. <http://ai.fon.bg.ac.rs/wp-content/uploads/2015/04/Classification-Decision-Trees-2015.pdf>.
- [34] S. K. Trivedi, "A study on credit scoring modeling with different feature selection and machine learning approaches," *Technol. Soc.*, vol. 63, no. September, p. 101413, 2020, doi: 10.1016/j.techsoc.2020.101413.
- [35] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 1–5, 2007, doi: 10.1016/j.eswa.2006.04.001.
- [36] J. Demšar et al., "Orange: Data Mining Toolbox in Python," *J. Mach. Learn. Res.*, vol. 14, pp. 2349–2353, 2013, [Online]. Available: <http://jmlr.org/papers/v14/demsar13a.html>.
- [37] IBM Corp., "IBM SPSS Statistics for Windows, Version 26.0," 2019. 2019.
- [38] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1999–2006, 2011, doi: 10.1016/j.eswa.2010.07.134.
- [39] S. Sayad, "An Introduction to Data Science," 2022. saedsayad.com/data_mining_map.htm.
- [40] IBM, "What is the k-nearest neighbors algorithm?," 2022. <https://www.ibm.com/topics/knn>.
- [41] O. Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, vol. 51. 2013.
- [42] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.
- [43] T. Yiu, "Understanding Random Forest," <https://towardsdatascience.com/>, 2019. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.
- [44] D. J. Livingstone, D. T. Manallack, and I. V. Tetko, "Data modelling with neural networks: Advantages and limitations," *J. Comput. Aided. Mol. Des.*, vol. 11, no. 2, pp. 135–142, 1997, doi: 10.1023/A:1008074223811.
- [45] S. A. Neslin, S. Gupta, W. Kamakura, L. U. Junxiang, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *J. Mark. Res.*, vol. 43, no. 2, pp. 204–211, 2006, doi: 10.1509/jmkr.43.2.204.