Sanches, Hugo Eduardo; Possebom, Ayslan Trevizan; Aylon, Linnyer Beatrys Ruiz

Article

Churn prediction for SaaS company with machine learning

Innovation & Management Review

Provided in Cooperation with:
University of São Paulo, School of Economics, Management, Accounting and Actuarial Sciences (FEA-USP)

Suggested Citation: Sanches, Hugo Eduardo; Possebom, Ayslan Trevizan; Aylon, Linnyer Beatrys Ruiz (2025) : Churn prediction for SaaS company with machine learning, Innovation & Management Review, ISSN 2515-8961, Emerald, Leeds, Vol. 22, Iss. 2, pp. 130-142, https://doi.org/10.1108/INMR-06-2023-0101

Received 20 June 2023
Revised 1 November 2023
30 April 2024
Accepted 17 June 2024

Churn prediction for SaaS company with machine learning

Hugo Eduardo Sanches
Department of Informatics, State University of Maringa - UEM, Maringa, Brazil
Ayslan Trevizan Possebom
Department of Informatics, Federal Institute of Parana - IFPR, Paranavai, Brazil, and
Linnyer Beatrys Ruiz Aylon
Department of Informatics, State University of Maringa - UEM, Maringa, Brazil

Abstract

Purpose – In an era marked by fierce business competition, customer retention is crucial for sustaining profitability. Churn prediction, the ability to forecast customer defections, is essential to enhance retention and can profoundly impact a company's bottom line. Among prediction techniques, machine learning techniques have proven to be efficient and reliable. Thus, this research aims to develop a model that effectively predicts customer churn for TecnoSpeed and provides insights into customer behavior.

Design/methodology/approach – Through a preprocessing and normalization of data, seven machine learning algorithms were applied. The models were trained, and also cross-validation and parameter tuning techniques were applied to improve results. The study also explores feature performance, providing insights into attributes that influence customer churn, thereby guiding effective strategies.

Findings – The results of three algorithms achieved over 90% accuracy, with less than 10% of the errors being part false negatives. We also introduce the Churn Probability Index, a novel metric that aggregates the outputs of multiple predictive models to provide an assessment of high-risk churn. This research is of significant importance as it contributes to the development of effective retention strategies for SaaS companies.

Originality/value – By applying machine learning to churn prediction, this study offers valuable insights into the performance and comparative analysis of different algorithms in a real-world SaaS environment. This study stands distinguished by its emphasis on a practical business scenario, enriched by a robust dataset provided and a large set of machine learning techniques. The findings provide practical implications for managers and administrators seeking to optimize customer retention and profitability.

Keywords Machine learning, Churn prediction, Data science, Software as a service

Paper type Research paper

1. Introduction

In the dynamic realm of the digital economy, Software as a Service (SaaS) platforms have revolutionized the way businesses operate. As these platforms burgeon in number and variety, SaaS companies face the challenge of retaining their customer base amidst fierce competition. Among the strategies of profit maximization, customer retention is the most effective and cost-efficient alternative [2]. Indeed, there is a consensus among business analysts and Customer Relationship Management (CRM) analysts that acquiring new customers is considerably more expensive than retaining existing ones. Therefore, these analysts argue that companies need to study the reasons for customer churn and the patterns behind their data [6]. Studying these

Funding: Thanks to @manna_team, the Araucária Foundation to Support Scientific and Technological Development of the State of Paraná (FA) and the National Council for Scientific and Technological Development (CNPq) - Brazil process 421548/2022-3 for the support.

Innovation & Management Review
Vol. 22 No. 2, 2025
pp. 130-142
Emerald Publishing Limited
2515-8961
DOI 10.1108/INMR-06-2023-0101

patterns is the key to anticipating and even predicting customer churn. Simultaneously, in the current world, a large amount of data is being generated by companies on a daily basis at a high growth rate, allowing the use of machine learning techniques that have already proven effective in making predictions in scenarios with a large volume of data (Kumar & Chandrakala, 2016).

Advancements in machine learning and artificial intelligence algorithms have significantly expanded the capability to predict customer attrition. Accurately forecasting churn is a pressing challenge for contemporary businesses, as churn prediction is essential for maintaining competitiveness in a global market and enhancing customer relationships (Geiler, Affeldt, & Nadif, 2022).

While the subject of churn prediction has seen considerable scholarly attention, the current landscape is punctuated with a diverse array of machine learning methodologies. Each offers its strengths and nuances, making the choice of an optimal method an intricate task. This paper, positioned at the intersection of theoretical robustness and practical relevance, aims to identify a model and strategies that effectively predict customer churn for Tecnospeed. Additionally, it seeks to provide insights into the behavioral patterns of the customers, thereby enhancing understanding and enabling targeted interventions. Through this dual focus, this study attempts to not only forecast churn but also to provide actionable intelligence that can inform more nuanced customer engagement strategies.

Moreover, one of our main contributions is in exploring this rich dataset, which contains prior client behavior traits that enable us to document new insights into the main determinants predicting future client churn. Through a nuanced examination, this research not only offers valuable insights into the mechanics of machine learning applied to churn prediction but also seeks to ground these insights in a real-world business context, thus enriching both the academic discourse and offering actionable insights for industry practitioners. By leveraging Artificial Intelligence techniques, particularly machine learning on datasets with service cancellation information, we hope to deduce customer behaviors and pinpoint potential churn candidates.

For this study, the data was processed and normalized in different ways, and then the following algorithms were applied: Naive Bayes, Decision Trees, Random Forest, Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Logistic Regression. The models were trained, and techniques such as cross-validation and parameter tuning were applied to improve the algorithm results. The SVM, Decision Trees, and Random Forest algorithms achieved accuracy rates exceeding 90%, while ANN and Logistic Regression achieved accuracies of over 88% and 87% respectively. Also, this paper introduces the Churn Probability Index (CPI) which represents an innovative approach that combines the predictive powers of multiple algorithms to generate a unified score indicative of churn risk. This index is designed to enhance the decision-making process by highlighting customers with a higher chance to churn.

Finally, an integral part of our study is the analysis of feature performance, which plays a crucial role in understanding the dynamics of customer churn. By identifying and evaluating the most influential features of some of our predictive models, we gain deeper insights into the factors that drive customer decisions to discontinue services. This analysis provides strategic guidance for businesses looking to implement targeted interventions. By pinpointing key attributes that influence churn, Tecnospeed and similar SaaS companies can allocate resources more effectively and design interventions that directly address the underlying causes of customer attrition.

2. Literature review

2.1 Churn

Companies must employ strategies that maximize profits, which can include acquiring new customers, selling more to existing customers, and most importantly, retaining their existing customer base. Like said in the introduction, It is widely recognized that acquiring new customers is often considerably more expensive than retaining existing ones (Lalwani, Mishra, Chadha, & Sethi, 2022).

Churn, often referred to as customer attrition or customer defection, is a crucial metric in the business world. It quantifies the rate at which customers discontinue their relationship with a company by discontinuing their use of its products or services (Manfrinatto, Striquer, & Wolf, 2020).

Understanding and managing churn is of paramount importance for companies across various industries. It serves as a critical gauge of customer dissatisfaction and operational issues and is instrumental in shaping customer retention strategies (Tékouabou, Gherghina, Toulni, Mata, & Martins, 2022).

2.2 Churn prediction models

Churn prediction models leverage historical customer data, encompassing a multitude of attributes and behaviors, to forecast which customers are at risk of churning in the future. These models typically employ machine learning algorithms to analyze patterns and identify factors associated with churn. By discerning the signals that precede customer defection, businesses can take proactive measures to retain valuable customers and prevent revenue erosion (Suh, 2023). Machine learning algorithms, have been deployed in churn prediction tasks especially across telecom and financial industries.

2.3 Business implications of churn prediction

Churn prediction carries profound implications for businesses. By harnessing the power of machine learning algorithms, companies can gain a competitive advantage in multiple ways. Firstly, the ability to foresee potential churners allows organizations to tailor retention strategies to individual customers, whether through personalized marketing campaigns or targeted interventions. Indeed, numerous studies show that preventing customer churn saves money, as acquiring new customers can cost up to five times as much as satisfying and retaining existing customers (De Lima Lemos, Silva, & Tabak, 2022).

Secondly, proactive churn prediction enhances the overall customer experience. By identifying and addressing the root causes of churn, businesses can make necessary improvements to their products, services, and customer support processes. This, in turn, leads to higher customer satisfaction, loyalty, and brand advocacy (Al-Najjar, Al-Rousan, & Al-Najjar, 2022).

In conclusion, churn prediction, powered by machine learning, is a critical tool in the modern business landscape. It empowers companies to mitigate customer attrition, maximize revenue, and foster enduring customer relationships.

3. Related works

This section will present some related works on the use of machine learning techniques in churn prediction. In a study conducted by Rahman and Kumar (2020), the authors explored customer behavior data from a bank to predict churn using machine learning techniques. The authors utilized the KNN, SVM, Decision Trees, and Random Forest algorithms for their research, with Random Forest algorithm achieving the highest accuracy of over 85% for the application.

In a similar research (Tékouabou *et al.*, 2022), the authors used six different machine learning algorithms to predict churn in a bank application with 12 different features obtaining 86% accuracy in the best score. In another bank application, the authors also used 6 different machine learning algorithms to a large Brazilian bank dataset with the best score found in Random Forest algorithm with 80% accuracy, the authors notice that this accuracy could save $80m USD in annual revenue (De Lima Lemos *et al.*, 2022).

In a different financial application, the authors applied five different machine learning algorithms to predict churn to credit card customers. The authors applied the machine learning models in a 12 month dataset and 16 different features and were able to predict with 90% accuracy with the C5 algorithm (Al-Najjar *et al.*, 2022).

In a different application, the authors used one machine learning technique to predict churn in the home appliance rental business field in Korea with a dataset of a known worldwide electronic company with 84,000 entries and 88% (Suh, 2023).

In another study by Lalwani *et al.*, 2022, the authors applied machine learning techniques to a telecommunications company's database using the Random Forest, Decision Trees, SVM, and Naive Bayes algorithms, achieving accuracy results above 80% in their model.

Further applications were found in the telecommunications industry, which was observed to be the most frequently studied application in machine learning research for churn prediction. According to different authors, the telecom industry is ideal for the utilization of machine learning techniques in churn prediction due to the large volume of data generated daily from a vast customer base (Qureshi, Arshad, Riaz, & Rasool, 2013; Vafeiadis, Kourentzes, Babai, & Chatzisavvas, 2015; Ullah *et al.*, 2019; Lalwani *et al.*, 2022).

In another similar study but with a different application, Rautio (2019) employed an approach based on Decision Trees, SVM, and Neural Networks to predict churn in a SaaS software company. The author used historical customer data to identify the most relevant factors for churn prediction and built a model that achieved an accuracy of over 85% accuracy.

Although several studies about this field were made, not only they are mostly focused on telecom and financial services churn, but also most of them apply just one to four ML techniques and an average of 15 features (Maan, & Maan, 2023). The present study explores a different real world business application with only a few studies in the literature, moreover we apply a large set of machine learning techniques and a more than 30 different features dataset.

4. Methodology

4.1 Dataset description

The datasets under scrutiny were graciously provided by Tecnospeed, a distinguished software company that specializes in Software as a Service (SaaS) solutions. These datasets, which form the bedrock of our research, offer valuable insights into customer churn dynamics, given the subscription-based nature of the company's offerings.

It is important to emphasize that stringent adherence to confidentiality agreements has been upheld throughout this research endeavor. Consequently, no personally identifiable information or sensitive corporate data has been divulged, preserving the privacy of both the company and its clients. All data utilized in this study has been meticulously anonymized to protect the identities of stakeholders involved.

Columns included various attributes. Among different datasets, key common attributes identified included: product, product type, average product ticket, average customer ticket, number of contracts for the same customer, number of chat interactions, number of call interactions, delinquency, lead source, normal customer support, urgent customer support, total customer support interactions, customer satisfaction survey results.

In addition to practical algorithm testing, correlation tests between attributes were conducted for a particular dataset. Some attributes such as chat interactions, call interactions, customer support interactions, and delinquency were separated into their quantities relative to different time periods: 0–3 months, 3–6 months, 6–9 months, and 9+ months. Customer satisfaction survey results were separated into good, bad, unanswered, and total results. Satisfaction surveys were categorized into positive, negative, unanswered, and total results. Data processing, algorithm implementation, and result analysis were all conducted using the Python programming language.

The dataset was reasonably balanced with 2,677 non-churn data rows and 2,234 churn data rows, representing 45.49% and 54.51% respectively comprising a total of 4,911 rows and 31 columns and is pertaining to the year 2021.

4.2 Data preparation

Initially, the data was subjected to the process of normalization and standardization, along with specific routines to handle missing, outlier, and categorical (string) data. Normalization and

standardization are performed to ensure that no attribute is assigned a higher weight in the algorithm than necessary due to a disproportionate value compared to the values of other attributes.

Lastly, handling categorical values is necessary to be able to use attributes with categorical values in different machine learning algorithms. In this treatment, categorical values were replaced with sets of Boolean values.

4.3 Training set and test set

Following the meticulous data preprocessing phase, the datasets were partitioned into training and test sets. The division adhered to the widely accepted 80–20 ratio, where 80% of the data constitutes the training set, and the remaining 20% serves as the test set. The data was also separated into attributes (X) and the class (Y), where the class is a Boolean variable indicating churn.

4.4 Parameter tuning and cross-validation

Parameter tuning emerges as an imperative step in our methodology. Different machine learning algorithms are equipped with unique parameters that demand calibration to align with the idiosyncrasies of the dataset and the overarching objectives of the research.

In the parameter tuning process, different parameters are tested for the dataset, and the best parameters are chosen for each algorithm. This process maximizes the accuracy of each algorithm and optimizes its performance for the given dataset.

For this research, we also employed a cross-validation algorithm encompassing 300 iterations, enabling a comprehensive evaluation of algorithm performance across diverse training and test sets. This approach not only bolsters the reliability of our results but also safeguards against the inadvertent inclusion of training data in the test set, averting data leakage.

The summary of the methodology for implementing and analyzing the algorithms can be found in Figure 1.

4.5 Performance evaluation

In this study, various performance metrics were meticulously assessed to provide a comprehensive evaluation of the model. These metrics included accuracy, precision, recall (sensitivity), F1-Score, and the ratios of False Positive (FP), True Positive (TP), False Negative (FN), and True Negative (TN) within the confusion matrix. To facilitate clarity in the subsequent calculations and discussions, the abbreviations FP, FN, TP, and TN will be employed to represent False Positive, False Negative, True Positive, and True Negative, respectively. The following formulas were utilized to compute the respective metrics:

$$\text{Accuracy} = (\text{VP} + \text{VN})/(\text{VP} + \text{VN} + \text{FP} + \text{FN}) \tag{1}$$

Accuracy, often referred to as the Positive Predictive Value, reflects the overarching effectiveness of the model, indicating the proportion of instances that were accurately classified.

$$\text{Precision} = \text{VP}/(\text{VP} + \text{FP}) \tag{2}$$

Precision elucidates the ratio of correctly predicted positive instances to all instances predicted as positive by the model.

$$\text{Recall} = \text{VP}/(\text{VP} + \text{FN}) \tag{3}$$

Also known as sensitivity, recall represents the proportion of actual positive instances that were correctly identified by the model.

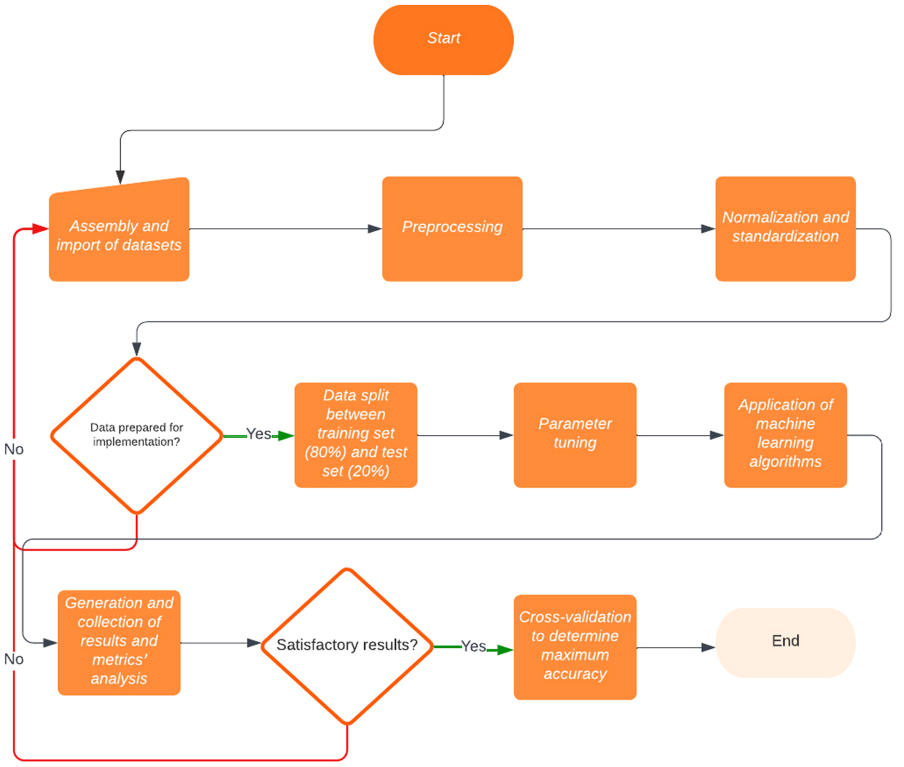

Figure 1. Flowchart of the methodology

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

The F1-Score is the harmonic mean of precision and recall.

While accuracy serves as a valuable preliminary indicator of model performance, it may yield deceptive results, particularly in imbalanced datasets. For instance, in a dataset where a mere 1% cancels the service/product (1% churn), a model predicting no cancellations will have 99% accuracy (and will generally be a useless model).

Precision, in a simple and straightforward way, can be seen as the percentage of correctly predicted items among all predicted items of a class. Precision is important in situations where false positives are more harmful than false negatives. For instance, when predicting whether a marketing campaign is good or not, the model needs to be assertive and identify only good marketing campaigns, even if it means miss classifying some good campaigns as bad.

Recall, in a simple and straightforward way, can be seen as the percentage of correctly predicted items among all instances of a particular class. Recall is used in situations where false negatives are more harmful than false positives. For instance, in a model that predicts customers likely to cancel a product or service, it is vital to capture as many churners as possible, even at the risk of falsely identifying some non-churners as potential churners.

In the scope of this research, miss classifying a customer who would not cancel as a customer who would cancel is not harmful (perhaps the company will only send one more marketing email to a customer who will simply ignore it). Therefore, since false negatives are more harmful than false positives, recall emerges as a particularly pertinent metric for evaluating the results of this study.

The F1-Score, given its nature as the harmonic mean between precision and recall, provides a general evaluation of the model. All metrics will be measured and analyzed separately and together to obtain the model with the best performance for the application.

4.6 Feature performance

Feature importance refers to a set of techniques aimed at assigning a score to input features based on how useful they are at predicting a target variable. In predictive modeling, particularly in problems involving high-dimensional data, not all features contribute equally to the accuracy of the model. Some features might be highly informative while others may be less irrelevant.

By identifying what factors most significantly impact the target variable, businesses can better allocate resources to address the most influential factors. For example, if feature importance analysis reveals that the frequency of service use is a top predictor of churn, a business might focus on strategies to increase user engagement.

Algorithms such as Decision Trees and Random Forests automatically compute feature importance during model training. The importance provided by these models is calculated based on how effectively each feature splits points into homogeneous sets.

For linear models as logistic regression, the coefficients associated with each feature can be considered as representing the feature's importance, provided that the data is standardized. In these models, a high absolute coefficient value indicates that a feature has a strong effect on the response variable, with the sign indicating the direction of the effect.

4.7 Churn probability index – CPI

The Churn Probability Index (CPI) is a composite metric designed to leverage the strengths of multiple predictive algorithms to estimate the probability of churn for each customer. By combining different algorithmic predictions, the CPI aims to mitigate the weaknesses of individual models and capitalize on their predictive power. This method is common in broader data science applications, such as credit scoring or disease risk assessment, but its adaptation for churn prediction offers a new avenue for business applications, particularly in enhancing customer retention strategies.

The CPI is designed to integrate predictions from multiple high-performing algorithms to estimate churn probability. Each selected algorithm contributes equally to the final index. This methodology assumes that each algorithm, having demonstrated high performance on its own, offers a unique perspective on the data, and that their equal integration can provide a more balanced and robust prediction. Also, the company is still able to verify which algorithm has made each vote so, it still has the decision power of following the best performance algorithm, but highlighting potential churns with unanimous votes. The selection of algorithms for inclusion in the CPI is based on their performance metrics from preliminary tests, including specially accuracy and recall.

Each algorithm is trained on the same training dataset and then used to predict churn on a validation set. The prediction from each algorithm is a binary output (0 for no churn, 1 for churn), representing the algorithm's "vote" for each customer's churn status (the number of chosen algorithms should be odd to avoid draw votes). The CPI for each customer is then calculated as the average of these votes. Mathematically, the CPI for a customer is expressed as:

$$\text{CPI} = \frac{1}{N} \sum_{i=1}^{N} p_i \tag{5}$$

where p_i is the prediction from the i-th algorithm, and N is the number of algorithms.

5. Results

The performance of each algorithm is evaluated using a suite of metrics and confusion matrices which provide a comprehensive view of how well each algorithm predicts churn and its associated operational implications. Accuracy was measured and will be presented in two ways: the basic accuracy of the model and the maximum accuracy generated after cross-validation.

In the following tables, 1 indicates churn, and 0 indicates no churn. The confusion matrices in Table 1 show the quantities of VP, VN, FP, and FN for each of the seven algorithms. In Table 2, the analysis metrics for each algorithm can be found.

In Table 3, we present the top four feature importance derived from Decision Trees and Random Forest, along with the top six coefficients from Logistic Regression. It is important to note that feature importance is represented as a percentage, with the sum of all importance equaling one, indicating a relative measure of each feature's contribution to the predictions of the model. Conversely, the coefficients from Logistic Regression can take any real value. Positive coefficients indicate attributes that are positively correlated with churn, suggesting that higher values of these attributes increase the likelihood of churn. Negative coefficients, on the other hand, signify attributes that are inversely related to churn, where higher values are associated with a decreased likelihood of churn. This distinction helps in understanding how different features influence the propensity for a customer to churn or remain with the service.

6. Discussion and analysis

In terms of accuracy, the decision tree, random forest, and SVM algorithms displayed superior performance, exceeding 90% accuracy, proving their efficacy in predicting churn in the given scenario. Both the logistic regression and ANN algorithms also yielded results close to 90% accuracy and could potentially be employed for the stated application.

The KNN and Naive Bayes algorithms performed below expectations, and the hypothesis of using the KNN algorithm for the application was discarded. Regarding the Naive Bayes algorithm, a result deserving attention was observed. Despite obtaining the worst accuracy among all the algorithms and performing worse than simply declaring all values as non-churn

Table 1. Confusion matrices of each algorithm

Naive Bayes	Real/Predicted	0	1
	0	147	403
	1	40	393
Decision trees	Real/Predicted	0	1
	0	521	29
	1	57	376
Random forest	Real/Predicted	0	1
	0	518	32
	1	64	369
KNN	Real/Predicted	0	1
	0	458	92
	1	139	294
Logistic regression	Real/Predicted	0	1
	0	502	48
	1	74	359
SVM	Real/Predicted	0	1
	0	523	27
	1	64	369
ANN	Real/Predicted	0	1
	0	501	49
	1	66	367

Table 2. Analysis metrics of the algorithms

Naive Bayes		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.79	0.27	0.4	0.549	0.59
	1	0.49	0.91	0.64		
Decision trees		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.9	0.95	0.92	0.9125	0.923
	1	0.93	0.87	0.9		
Random forest		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.89	0.94	0.92	0.902	0.906
	1	0.92	0.85	0.88		
KNN		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.77	0.83	0.8	0.765	0.765
	1	0.76	0.68	0.72		
Logistic regression		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.87	0.91	0.89	0.8759	0.88
	1	0.88	0.83	0.85		
SVM		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.89	0.95	0.92	0.907	0.907
	1	0.93	0.85	0.89		
ANN		Precision	Recall	F1-Score	Accuracy	Accuracy with Cross Validation
	0	0.88	0.92	0.9	0.885	0.886
	1	0.89	0.85	0.87		

Table 3. Top feature importance

Decision tree	Importance	Random forest	Importance	Logistic regression	Coefficients
id_product	0.09493898	id_product	0.07791878	satisfaction_good	−1.20999373
age_months	0.13356037	age_months	0.10381056	ticket_product_9-12	0.953659487
ticket_product_3	0.09676527	ticket_product_3	0.11560972	age_months	−0.964291186
ticket_product_9-12	0.17110625	ticket_product_9–12	0.09087104	ticket_customer_3	−0.716691044
Total	0.49637087	Total	0.3882101	satisfaction_bad	0.648694843
				ticket_product_3	−0.64078647

(0), the algorithm had the lowest number of FN among all the algorithms, and therefore, it had the highest recall for churn (1) compared to the others, reaching 91% (meaning that the algorithm correctly predicted 91% of the users who would churn). This

Shows that, despite performing well below average, the algorithm can still be used in some cases where the company is not concerned about the possibility of false positives (erroneously identifying a user who would not churn as a user who would churn).

This strategy can be useful in certain segments where anti-churn measures do not run the risk of "irritating" the customer with marketing actions that may be considered spam and drive the customer away instead of retaining them. However, correctly identifying a larger number of churners can bring competitive advantages to the company depending on its strategy.

Overall, the Decision tree algorithm is recommended for this application when appropriately tuned, as it outperforms other methods by approximately 1% to 2% compared to SVM and random forest (which ranked second and third). Moreover, it recorded the second-lowest FN count, trailing only the Naive Bayes algorithm, and therefore achieved the second-highest recall for churn (1) at 87%, roughly 2% higher than SVM and random forest.

Therefore, the use of the Random Forest, SVM, ANN, and Logistic Regression algorithms should not be ruled out, as changes in the dataset may yield better results with these algorithms and all the five of them (including Decision Trees) will be used for CPI's votes. For this specific business scenario, it is recommended to always test the top five algorithms before

choosing the ideal one for churn prediction, or to use them combined as in CPI approach. Naive Bayes can be used only when the company employs strategies that do not care about false positives. The use of the KNN algorithm is not recommended in this specific business scenario.

The objective of the Churn Probability Index (CPI) is not to enhance the performance of individual algorithms or the overall predictive system, but rather to identify and highlight high-potential churners, particularly in cases of unanimous votes from all selected algorithms. By focusing on instances where all algorithms agree on a churn prediction, the CPI aims to pinpoint customers at the highest risk of churn with a higher degree of confidence. This method assumes that when all algorithms, each with different underlying mechanics and sensitivities, converge on the same prediction, the likelihood of an accurate churn forecast is significantly increased. This approach enables businesses to prioritize and tailor their intervention strategies more effectively, directing resources and attention to those customers whose churn prediction is most certain, thereby optimizing retention efforts and potentially increasing the impact of targeted customer retention programs.

In the feature importance analysis, the same four attributes were prominently highlighted by both Decision Trees and Random Forest algorithms. Specifically, Decision Trees attributed nearly 50% of its decision-making to these attributes, while for Random Forest, these attributes accounted for approximately 40% of its decisions. This convergence provides critical insights into the decision-making process, which will be further explored in subsequent sections. Additionally, the Logistic Regression model identified three attributes that overlap with those highlighted by the Decision Tree and Random Forest models but presented a distinct perspective on their influence. The implications of these findings, including how they diverge across models, will also be discussed in detail below.

In conclusion, it is imperative to underscore the fact that the trained model, demonstrating commendable levels of accuracy and recall, has garnered approval from the company. Subsequently, it has been actively utilized by Tecnospeed for the systematic identification of churners throughout the duration of 2023. Also, the CPI approach has been approved and utilized by TecnoSpeed since 2024.

6.1 Business implications

While this research yielded significant insights into the realm of churn prediction using machine learning algorithms, it's imperative to consider its practical implications in a real-world business context, particularly for SaaS companies. While machine learning provides robust methodologies for predictions, their application in the dynamic business environment demands flexibility and adaptability.

Moreover, it's essential to iterate that while Decision Trees emerged as the top-performing algorithm in this specific study, the ideal algorithm can vary based on the unique characteristics and nuances of each dataset. Thus, it's recommended for businesses to trial multiple algorithms tailored to their datasets before settling on an optimal solution.

Conversely, it's crucial to underscore the significance of even a 1%-2% variation in accuracy and recall in practical business contexts. To put this into perspective, consider that Tecnospeed garnered just over ten thousand new signatures in 2021. A 2% discrepancy in recall implies that 200 potential churners weren't appropriately flagged prior to actual churn.

Such observations emphasize the need to continually strive for the optimal algorithm tailored to specific applications – additionally, it shows that the significance of methodologies such as cross-validation and parameter tuning cannot be overstated. While the majority of studies referenced in the related works section reported accuracies below 90% (with a substantial number hovering around 85%), a 5% difference cannot be dismissed as insignificant.

In the context of churn prediction for Tecnospeed, the analysis of feature importance and model coefficients provides critical insights for strategic decision-making. For instance, the

age of the contract emerged as a significant factor across all tree-based algorithms, where a longer contract duration correlates negatively with churn likelihood. This inverse relationship, as indicated by the negative regression coefficients, suggests that customers with extended contracts are less prone to churn.

Further, the importance of product characteristics was highlighted by Decision Tree and Random Forest models, which identified the Product ID as a crucial predictor of churn. These findings indicate that certain products are more susceptible to churn than others. Similarly, customer satisfaction plays a pivotal role, with Logistic Regression coefficients revealing that high customer satisfaction significantly reduces the likelihood of churn. Interestingly, while poor satisfaction increases churn risk, its effect is comparatively half as potent as the protective effect of high satisfaction, underscoring that satisfied customers are significantly more likely to remain loyal than dissatisfied customers are to leave.

Moreover, all tree-based algorithms consistently pointed to the average ticket price over the year as a key predictor, with higher-priced contracts more likely to be terminated. This trend was most pronounced in the outputs from Decision Trees and Random Forest. However, the analysis also showed that the initial pricing during the first three months (the implementation phase) holds substantial weight in churn predictions, being the second most influential factor for Decision Trees and Random Forest. Logistic Regression analysis further illustrates a contrasting effect, where a higher initial cost is associated with a lower churn probability. This nuanced finding suggests that products with higher upfront costs but lower ongoing expenses are less likely to experience churn.

These insights are particularly relevant for companies like Tecnospeed, where adjusting the pricing strategy—increasing implementation costs while reducing ongoing maintenance fees—may effectively lower the overall churn risk. Such strategic adjustments can help maintain a competitive average ticket price over twelve months, aligning with the company's broader customer retention objectives.

Furthermore, it's apparent that a company's capability to devise retention strategies for identified churners is equally vital. Consequently, pinpointing the root cause of the churn becomes as crucial as identifying the churner itself. In light of this, a promising avenue for future research is the integration of Explainable Artificial Intelligence, which could elucidate the reasons behind churn, thereby providing other invaluable insights for strategic decision-making.

7. Conclusion

Given the increasingly fierce competition that most companies face today and the rising cost of customer acquisition, it is crucial for companies to be attentive to customer retention issues and try to mitigate churn as quickly as possible. Therefore, predicting churn is necessary to establish customer retention strategies and maintain competitiveness in the current market (Lalwani *et al.*, 2022). Our research ventured into this realm, exploring the development of a model to effectively predict churn and also to provide insights of customer behavior within a specific SaaS company.

Our findings underscore that while all the seven applied algorithms have potential, the Decision Tree algorithm, in the provided dataset scenario, manifested superior performance. However, given the ever-evolving nature of datasets and the intricate dynamics of the business world, this superiority might not be universally applicable. For instance, while the Random Forest, SVM, ANN, and Logistic Regression algorithms trailed closely, further dataset modifications might alter their respective efficiencies.

It is always recommended to test the top five algorithms before choosing the best one for churn prediction, or to use them combined as in the CPI strategy (which is applied in the company nowadays), taking into account the reality of the dataset within the specified period. The Churn Probability Index (CPI) developed in this study has been applied as a valuable tool for assessing high rick churn. By synthesizing outputs from multiple predictive models, the

CPI offers a robust and nuanced metric that businesses can leverage to identify churn high risk customers more effectively.

Moreover, our investigation into feature performance has yielded significant insights into the key reasons of customer churn. By identifying which attributes most strongly influence churn decisions, we have provided a foundational understanding that can inform more effective customer retention strategies for Tecnospeed. This analytical approach empowers businesses to focus their efforts on modifying or bolstering the factors that most impact customer retention. Our research, while aligned in terms of methodologies and algorithms with similar works (with a larger set of algorithms), offers a unique value proposition by integrating the business context more cohesively.

For future work, the use of argumentation systems techniques to justify the results obtained by machine learning techniques is suggested. This approach would facilitate the identification of the underlying causes of customer churn, thereby enhancing the company's ability to develop and implement effective retention strategies for those identified as likely to churn. This would also make the results more explainable and provide other important data for decision-making.

References

Al-Najjar, D., Al-Rousan, N., & Al-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), 1529–1542. doi: 10.3390/jtaer17040077.

De Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing & Applications*, *34*(14), 11751–11768. doi: 10.1007/s00521-022-07067-x.

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, *14*(3), 217–242. doi: 10.1007/s41060-022-00312-5.

Kumar, A. S., & Chandrakala, D. (2016). A survey on customer churn prediction using machine learning techniques. *International Journal of Computer Applications*, *975*, 8887.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, *104*(2), 271–294. doi: 10.1007/s00607-021-00908-y.

Maan, J., & Maan, H. (2023). Customer churn prediction model using explainable machine learning. *International Journal of Computer Science Trends and Technology (IJCST)*, *11*(1), 34–40.

Manfrinatto, G. R., Striquer, L. P. S., & Wolf, A. S. (2020). Análise e Controle do Crescimento de Startups. *Caderno PAIC*, *21*(1), 97–112.

Qureshi, S. A., Arshad, J., Riaz, M. M., & Rasool, R. (2013). Telecommunication subscribers' churn prediction model using machine learning. In *Eighth international conference on digital information management (ICDIM 2013)* (pp. 131–136). IEEE.

Rahman, M., & Kumar, V. (2020). Machine learning based customer churn prediction in banking. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1196–1201). IEEE.

Rautio, A. J. O. (2019). Churn prediction in SaaS using machine learning. Dissertation. Faculty of Management and Business, Tampere University.

Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. *Journal of Big Data*, *10*(1), 41. doi: 10.1186/s40537-023-00721-8.

Tékouabou, S. C. K., Gherghina, Ş. C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. *Mathematics*, *10*(14), 2379. doi: 10.3390/math10142379.

Ullah, I., Mahmood, A., Shamsi, H. S., Baig, A. R., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn

prediction and factor identification in telecom sector. *IEEE Access*, *7*, 60134–60149. doi: 10.1109/access.2019.2914999.

Vafeiadis, T., Kourentzes, N., Babai, M. Z., & Chatzisavvas, K. (2015). A comparison of machine learn- ing techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. doi: 10.1016/j.simpat.2015.03.003.

Further reading

Abiodun, O. I., Azeez, S. A., Adeyemo, T. A., Alaba, F. A., Dada, K. V., Umar, A. M., . . . & Gana, U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*, *7*, 158820–158846. doi: 10.1109/access.2019.2945545.

Corresponding author
Hugo Eduardo Sanches can be contacted at: hugoe.sanches@gmail.com

Associate Editor: Ana Lucia Figueiredo Facin