Customer Churn Prediction Methods: Analysis and Evaluation

Gabriela Gjorgievska[1], Riste Stojanov[1], Gjorgjina Cenikj[3], Tome Eftimov[3], and Dimitar Trajanov[1,2]

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Republic of N. Macedonia
gabriela.gjorgievska@students.finki.ukim.mk,
{riste.stojanov,dimitar.trajanov}@finki.ukim.mk
[2] Computer Science Department, Metropolitan College, Boston University, Boston, USA
{dtrajano}@bu.edu
[3] Computer Systems Department, Jožef Stefan Institute, Ljubljana, Slovenia
{tome.eftimov,gjorgjina.cenikj}@ijs.si

Abstract. Customer retention is one of the primary pillars of product growth with the subscription-based business model. The competition is fierce in the SaaS (Software as a service) market, where customers are free to choose from the many companies that offer similar competitive services. Sometimes, multiple bad experiences, or even a single one, can make the customer give up on the product or on the service. Globalization and rising competition are increasing the cost of getting a new customer, making it more affordable to invest in retaining customers. Therefore, it is crucial for any business to be able to anticipate its customers' behavior and try to prevent the end of their cooperation, i.e., to predict the churn of customers.

Keywords: Churn prediction · Machine learning · Model evaluation · XGBoost.

1 Introduction

The percentage of customers that stop interacting with a given business is called churn rate, while the process of identifying those customers is called churn prediction. Nowadays, churn prediction and management are critical for more and more companies in the fast-changing, intensely competitive, and saturated market. To improve customer retention, companies must be able to predict customers at risk who are prone to switch service providers or products for the competition. Therefore, many companies are adopting data mining techniques for churn prediction and analysis. An effective and accurate customer churn prediction model allows companies to target customers that are most likely to abandon them (churn) and use various marketing strategies to retain those customers. Customer retention is profitable for companies since attracting new customers is five to six times more expensive than retaining already existing customers [21].

Currently, it is a trend to collect a huge amount of customer data, which can enable the recognition of behavioral patterns of potential churning customers. This knowledge can help in segmenting customers and thus take measures to retain them. For these purposes, many studies have been conducted [1][7][19][9], showing that using machine learning models can help accurately predict the churners. However, collecting huge amounts of data comes with a cost, and collecting more data increases the price of storing and processing this data. However, not all collected information is equally valuable, and some features improve the churn prediction, while others may not influence it at all.

The remainder of this paper is organized as follows. In Section 2, the literature related to data mining techniques for customer churn is reviewed. Section 3 describes the process of data acquisition and the datasets being used. In Section 4, the data pre-processing is described. Section 5 presents the proposed models that were used for churn prediction. Section 6 contains experimental results. Finally, the conclusion is provided in Section 7.

2 Related Work

A number of studies using different algorithms and techniques have been conducted to explore the customer churn problem, as well as the possibilities and potential of providing solutions using data mining techniques.

In [1], the authors approached the problem with multiple techniques including Decision Trees (DT), Random Forest (RF) [16], Gradient Boosted Machine Tree (GBM) [15] and Extreme Gradient Boosting (XGBoost) [6]. The models' performance is evaluated using the Area Under the Curve (AUC) measures on a dataset of 5 million customers, with a result of 93% on the training data set and 89% on the testing dataset.

In [7], logistic regression, decision trees, GBM, and k-nearest-neighbors(KNN) are some of the methods used to predict the churning of customers on a dataset containing 640,000 customers with 65,000 churners. The best result is 89% F1-score, obtained with the GBM method.

Authors of study [19] analyze backpropagation network, support vector machines (SVM), decision trees, Naive Bayes, and logistic regression with boosting (AdaBoost and M1 algorithms) and without boosting. The authors conclude that using a boosting algorithm improves the accuracy of the model up to 4% and the F1-score up to 15%.

In [9], the authors use the logit leaf model, which is actually a hybrid model that uses decision trees and logistic regression. The decision tree is used to separate the customers into homogeneous segments, and then on these segments, the logistic regression model is used. The authors also used a random undersampling technique for dataset balancing.

Authors of [13] use logistic regression and logit boost models for customer churn prediction problems with F1-score from 80% to 81%.

In [20] is used an improved balanced RF model, which is a combination of balanced and weighted RF. Their dataset contains 20,000 customers with 27 attributes, and they have obtained an accuracy of 93%.

RF and KNN models are used in [12]. PSO sampling and random undersampling techniques are used for dataset balancing. mRMR (Minimum redundancy maximum relevance) technique is used for feature selection, which selects the attributes that have a strong correlation with the dependent variable but minimal redundancy with variables that were chosen before. The RF model gives the best result with 75% AUC.

In [4], an SVM classifier is used to predict churning customers. The dataset contains 3,333 customers with 21 features. The dataset is partitioned 80% for training and 20% of the data for testing the classifier. Because SVM requires the classes to be balanced, the authors use boosting for balancing. The end result is 88% of accuracy using the polynomial kernel function, although all three models (using polynomial, rbf, or linear kernel functions) perform really well. The worst performance was using the sigmoid function.

In [18], decision trees and neural networks are used for churn prediction. Accuracy with the DT model was 98.88%, and the neural network had an accuracy of 98.43%. The dataset had 18,000 customers with 252 attributes for each customer.

In [11], authors use a decision tree and backpropagation network with k-means. Authors used call details and billing information to train their models. The dataset used in this study contains 160,000 customers. The imbalanced dataset problem was resolved with oversampling.

The authors of [14] have collected a number of different papers that try to solve the customer churn problem and made a survey of all methods in order to find the best solution for customer churn prediction. They identified that the most popular methodology is the neural network classifier, and a close second is the DT classifier. A survey of the most popular papers to date is published in [2]. They summarize and describe the existing public datasets. Many of the papers use decision tree algorithms as the classifier of churn as well as logistic regression and artificial neural network classifiers. The most commonly used performance metrics are the confusion matrix, F1 score, and AUC. In [8] and [3], authors used the decision trees model on telecommunication datasets.

3 Data Acquisition

For the purpose of this paper, 10 datasets were used. These datasets are mainly from the telecommunication industry since the churn impact was first identified there. For data acquisition, we used the following sources: kaggle.com[4], data.world[5], and google data sets[6], with "churn prediction" as a search query. In the following text the details about the selected datasets are presented:

[4] https://www.kaggle.com/search?q=churn+prediction
[5] https://data.world/
[6] https://datasetsearch.research.google.com/search?query=churn%20prediction

1. **Customer Churn Prediction dataset**: This dataset was found on kaggle[7]. It contains 4,250 samples for the training set and 750 samples in the testing set. It is a dataset from the telecom industry containing features like the number of customer service calls made by the customer, total daily charges, minutes, and calls made by the customer if the customer has a voicemail plan, the number of voicemail messages, how much time the customer has the account and so on. The dataset has 20 features for each customer.
2. **Network provider customer churn**: This dataset was found on kaggle[8]. The dataset is from a network provider, it has 7,043 samples of data, each with 21 features. Some of the features are monthly charges, total charges, payment method, if the customer has activated paperless billing, if the customer has included streaming TV or movies, if the customer has activated multiple lines, online security, online backup option, etc.
3. **Bank churners**: The Bank churners dataset was found on kaggle[9]. It is a bank dataset containing 10,127 samples of data. For each customer, the following information is provided: customer age, gender, education level, marital status, customer income, credit limit, number of inactive months, and so on. It has 23 features per sample.
4. **Telecom churn**: This dataset was found on kaggle[10]. The dataset contains 3,333 customers, for each customer dataset has 11 features, including roaming minutes, monthly charges, daily calls, daily minutes, customer service calls, is the contract renewed, etc.
5. **Bank Customer Churn**: The Bank customer churn dataset was found on kaggle[11]. It is a dataset for bank customers, it has e 5,000 samples with 15 features. For each customer the bank keeps the following information: customer credit score, gender, age, if the customer has a credit card, if it is an active member, estimated salary, customer balance, and so on.
6. **UCI Churn dataset**: This dataset was found on data.world[12]. It has e 5,000 samples of data, with a total of 18 features. This dataset has been anonymized, in order to protect sensitive customer information. It contains features like the number of phone calls made to customer service, if the user has a voicemail plan, how much time the customer has been a subscriber, number of daily minutes, number of daily calls, if the user has an international plan, number of minutes or calls made using the international plan, etc.
7. **South Asian Churn dataset**: South Asian Churn dataset was found on kaggle[13]. The data set used is real-life data collected from a major wireless telecom operator in South Asia. It has 15 features and 2000 samples. For each customer in the dataset, the company keeps the following features: total

[7] https://www.kaggle.com/c/customer-churn-prediction-2020/data
[8] https://www.kaggle.com/lokeshkum/network-provider-customer-churn-data
[9] https://www.kaggle.com/syviaw/bankchurners
[10] https://www.kaggle.com/barun2104/telecom-churn
[11] https://www.kaggle.com/santoshd3/bank-customers
[12] https://data.world/earino/churn
[13] https://www.kaggle.com/mahreen/sato2015

monthly revenue, total sms revenue, total data revenue, how many years the customer has been a subscriber, other favourite network, type of the user and so on.

8. **Customer churn**: This dataset was found on data.world[14]. The dataset is from the telecom industry and it contains 21 features. It has 2,998 samples, and for each customer, the dataset has the following features: total call minutes, number of calls and total charge throughout the day, the evening and night, if the customer has an international plan included, if the customer has spent international minutes, number of calls and the total charge, number of calls made to customer support, etc.
9. **Telecom customer churn**: This dataset was found on kaggle[15]. The dataset originates from the telecom industry. It contains 1,401 samples with a total of 16 features, including the churn feature. For each customer the company saves information like the total amount of data spent, the total number of sent sms, the total number of unique calls, the number of complaints made by the customer, etc.
10. **Binary Customer churn**: This dataset was found on kaggle[16]. It contains 900 samples, with a total of 10 features, including the churn feature. This dataset is from a marketing company that is trying to determine the churner clients. For each client, the dataset contains the client's name, age, number of total advertisements purchased, total years of being a client, number of websites that use the service, client address, name of the company, etc.

4 Data Pre-processing

In this section, all of the pre-processing techniques performed on the datasets are described in more detail so that the dataset will be ready for the model training. Each dataset was checked for null values for all features, so any feature that contained more null values than actual values was removed. Additionally, all data samples (i.e., instances) that contain more null values than actual values were removed, and all categorical features were processed and encoded.

For all datasets, we used stratified training to test split with a ratio of 70% to 30%. The hyperparameters of the algorithms were optimized using repeated stratified K-fold cross-validation, using 10 folds and 3 repeats. The datasets that we extracted are unbalanced, and this could cause a significant negative impact on the models being trained with them. This problem was resolved by balancing the sample of training data by selecting the samples such that the two classes were balanced. The SMOTE (Synthetic Minority Oversampling TEchnique) [5] technique was used in order to select the training samples.

[14] https://data.world/marktmilligan/customerchurn
[15] https://www.kaggle.com/filemide/churns
[16] https://www.kaggle.com/hassanamin/customer-churn

Table 1. F1 score for each algorithm

Dataset	Logistic Reg.	Decision tree	Random forest	SVM	Gradient Boosting	XGBoost	MLP classifier	Neural Network
1	50	72	69	67	67	**80**	66	71
2	**61**	58	53	60	58	**61**	**61**	**61**
3	89	96	96	95	**97**	**97**	93	**97**
4	47	70	67	62	61	**72**	58	61
5	46	**57**	55	56	56	**57**	55	56
6	47	73	75	60	67	**79**	59	66
7	73	69	66	**75**	69	72	72	74
8	46	**76**	61	60	63	75	60	66
9	66	70	74	73	**78**	**78**	69	68
10	60	64	60	57	62	54	63	**65**

5 Proposed Models

In this section are described the classifiers that were trained on all previously collected datasets. We identified the most frequent and most successful algorithms analyzed in Section 2. Here are the details about the parameters of each algorithm:

5.1 Logistic regression

The first classifier, logistic regression (LR), is considered a baseline classifier as the reference for all other classifiers. LR is a transformation of linear regression using the sigmoid function, which is a process of modeling the probability of a discrete outcome given an input variable. For all data sets, LR has been used with parameter C and values 1, 5, 10.

5.2 Decision tree

Decision tree (DT) is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For all data sets, DT classifiers have been used with criteria that measure the quality of the split with value entropy, maximum depth of the tree 7, and a minimum number of samples required for a node to be a leaf node with value 30.

5.3 Random forest

Random forest (RF) [16], as its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the RF makes a class prediction, and the class with the most votes becomes our model's prediction. In this study, an RF classifier is used with parameter

n_estimators and values of 1, 5, and 10, which represent the number of trees in the forest.

5.4 Support vector machine

Support Vector Machine (SVM) [10] is one of the most popular supervised learning algorithms. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that the new sample of data can be easily put in the correct category. This best decision boundary is called a hyperplane. SVM chooses the extreme vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed a support vector machine. In this study SVM algorithm is used with the following parameters gamma with value auto, C with values 10, and 15, kernel with values rbf and linear, which represents the kernel type that is used in the algorithm.

5.5 Gradient boosting machine

Gradient Boosting Machine (GBM) [15] is an iterative functional gradient algorithm, which minimizes a loss function by iteratively choosing a function that points towards the negative gradient, a weak hypothesis. GBM produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. In this study, GBM is used with n_estimators 20, the learning rate of 0.75, maximum features of 4, and a maximum depth of 5.

5.6 XGBoost

XGBoost [6] stands for eXtreme Gradient Boosting. It is a DT-based ensemble machine-learning algorithm that uses a gradient-boosting framework. XGBoost is an implementation of gradient-boosted DTs designed for speed and performance. It is used with the following parameters, maximum depth of trees of 7 and learning rate of 0.5.

5.7 Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) [17] is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers, and each layer is fully connected to the following one. The nodes of the layers are neurons using nonlinear activation functions, except for the nodes of the input layer. There can be one or more non-linear hidden layers between the input and the output layer. In this study, the MPL classifier is used with the following parameters activation with value relu (rectified linear unit function), solver sgd (stochastic gradient descent), and learning rate adaptive.

5.8 Artificial neural network

And last but not least, there is also an artificial neural network model. For this model, Keras Sequential models are used. A 64–8–1 dense, layered model with a decaying learning rate of batch size 32 is used. L2 regularization and dropout for each layer are also used. ANNs consist of an artificial network of functions called parameters, which allows the model to learn and fine-tune itself by analyzing new data. Each parameter is a function that produces an output after receiving one or multiple inputs. Those outputs are then passed to the next layer of neurons, which use them as inputs of their own function, and produce further outputs, and so it continues until every layer of neurons has been considered and the terminal neurons have received their input. Those terminal neurons then output the final result for the model.

6 Evaluation and Discussion

The results of the churn prediction experiments are of utmost importance in this study as they provide a clear picture of the performance of different machine learning algorithms in predicting customer churn. We created 80 distinct customer churn prediction models by utilizing ten datasets we collected and eight machine learning models we selected. Table 1 provides a comprehensive summary of the results, with the F1 score of each classifier on each dataset displayed. The F1 score is a widely used metric for evaluating the performance of binary classification models and takes into account both precision and recall.

One of the key findings from Table 1 is the identification of the top-performing models for each dataset, as represented by the bolded numbers. The results show that the XGBoost classifier emerged as the best classifier for seven out of the ten datasets, with the artificial neural network classifier following closely as the second-best option in many cases.

XGBoost's ability to handle larger datasets and its robustness to imbalanced datasets likely contributed to its performance in the majority of the cases. While XGBoost performed well in most cases, it's still important to consider the trade-off between accuracy and interpretability when choosing a model. XGBoost is known for its accuracy, but its complexity can make it difficult to understand and interpret the model's predictions. In contrast, Decision Trees and SVM are often simpler and more interpretable, making it easier to understand the factors driving customer churn.

In addition to the F1 scores, Table 2 presents the confusion matrix values for each dataset for the best classifier. In the confusion matrix, the values for true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are presented. Looking at these values, we can get a comprehensive overview of the model's performance, enabling a thorough evaluation of its strengths and weaknesses.

The results highlight the significance of carefully evaluating and selecting models to ensure accurate and effective predictions. The experiments also underline the need for considering multiple algorithms and assessing their performance

Table 2. Confusion matrix values for each dataset for the best classifier

Dataset	Algorithm	TP	TN	FP	FN
1	XGBoost	1053	149	32	32
2	XGBoost	1268	374	225	225
3	XGBoost	456	2460	40	40
4	XGBoost	792	119	50	50
5	XGBoost	1968	420	416	416
6	XGBoost	1214	189	56	56
7	SVM	201	242	101	101
8	DT	729	106	30	30
9	XGBoost	176	158	38	38
10	ANN	205	31	12	22

on various datasets to ensure the best possible results. This study sheds light on the importance of using machine learning algorithms for churn prediction and provides valuable insights for businesses looking to implement such models.

7 Conclusion

Churn prediction is the process of identifying customers who are likely to stop using a company's products or services in the near future. This is a valuable tool for businesses because it allows them to proactively address customer satisfaction and loyalty issues and potentially retain customers who would otherwise leave. One of the most efficient ways for churn prediction is through the use of machine learning algorithms, which analyze past customer behavior and other relevant data to identify patterns and make predictions.

In this paper, first, we collected and analyzed ten datasets for churn prediction. The size of the datasets varied greatly, starting from a couple of hundred to a couple of thousand samples, and they included features such as demographic information, usage data, customer service interactions, and feedback.

Using the collected ten datasets and selected eight machine learning models, we developed 80 different churn prediction models. The analysis and evaluation showed that XGBoost was the best-performing model in seven cases. This highlights the strength and versatility of the XGBoost algorithm in predicting customer churn. On the other hand, SVM, Decision Trees, and Neural Networks were the best performers in one case each, demonstrating their potential as effective models for churn prediction.

These results demonstrate the importance of evaluating multiple machine-learning models and datasets when predicting customer churn. No single algorithm or model is guaranteed to perform the best in all cases, and it is important to consider the specific characteristics and requirements of each business and dataset when choosing a churn prediction model.

Overall, the results of this study provide valuable insights for businesses looking to implement churn prediction models and highlight the importance

of carefully selecting and evaluating models to ensure accurate and effective predictions.

Ultimately, it's important to note that churn prediction is not an exact science and that the accuracy of the model can vary depending on the quality and quantity of data available, as well as the choice of algorithm and features used. However, even imperfect churn prediction models can still provide valuable insights and help businesses reduce churn and improve customer satisfaction.

Acknowledgment

This work is supported by the Slovenian Research Agency (research core funding No. P2-0098).

This work was also supported by the NLP4Food (Use of NLP in the food domain) project funded by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje, N. Macedonia

References

1. Ahmad, A., Jafar, A., Aljoumaa, K.: Customer churn prediction in telecom using machine learning in big data platform. j. big data 6 (1), 1–24 (2019)
2. Ahn, J., Hwang, J., Kim, D., Choi, H., Kang, S.: A survey on churn analysis in various business domains. IEEE Access **8**, 220816–220839 (2020)
3. Bin, L., Peiji, S., Juan, L.: Customer churn prediction based on the decision tree in personal handyphone system service. In: 2007 International Conference on Service Systems and Service Management. pp. 1–5. IEEE (2007)
4. Brandusoiu, I., Toderean, G.: Churn prediction in the telecommunications sector using support vector machines. Margin **1**, x1 (2013)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al.: Xgboost: extreme gradient boosting. R package version 0.4-2 **1**(4), 1–4 (2015)
7. Chouiekh, A., et al.: Machine learning techniques applied to prepaid subscribers: case study on the telecom industry of morocco. In: 2017 Intelligent Systems and Computer Vision (ISCV). pp. 1–8. IEEE (2017)
8. Dahiya, K., Bhatia, S.: Customer churn analysis in telecom industry. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions). pp. 1–6. IEEE (2015)
9. De Caigny, A., Coussement, K., De Bock, K.W.: A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research **269**(2), 760–772 (2018)
10. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)
11. Hung, S.Y., Yen, D.C., Wang, H.Y.: Applying data mining to telecom churn management. Expert Systems with Applications **31**(3), 515–524 (2006)
12. Idris, A., Rizwan, M., Khan, A.: Churn prediction in telecom using random forest and pso based data balancing in combination with various feature selection strategies. Computers & Electrical Engineering **38**(6), 1808–1819 (2012)

13. Jain, H., Khunteta, A., Srivastava, S.: Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science **167**, 101–112 (2020)
14. Kamalraj, N., Malathi, A.: A survey on churn prediction techniques in communication sector. International Journal of Computer Applications **64**(5), 39–42 (2013)
15. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers in neurorobotics **7**, 21 (2013)
16. Pal, M.: Random forest classifier for remote sensing classification. International journal of remote sensing **26**(1), 217–222 (2005)
17. Riedmiller, M., Lernen, A.: Multi layer perceptron. Machine Learning Lab Special Lecture, University of Freiburg pp. 7–24 (2014)
18. Umayaparvathi, V., Iyakutti, K.: Applications of data mining techniques in telecom churn prediction. International Journal of Computer Applications **42**(20), 5–9 (2012)
19. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.C.: A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory **55**, 1–9 (2015)
20. Xie, Y., Li, X., Ngai, E., Ying, W.: Customer churn prediction using improved balanced random forests. Expert Systems with Applications **36**(3), 5445–5449 (2009)
21. Zhu, B., Baesens, B., vanden Broucke, S.K.: An empirical comparison of techniques for the class imbalance problem in churn prediction. Information sciences **408**, 84–99 (2017)