







# Machine Learning for Customer Churn Prediction in Retail Banking

Joana Dias<sup>1,2</sup> , Pedro Godinho<sup>2</sup>  , and Pedro Torres<sup>2</sup> 

<sup>1</sup> INESC-Coimbra, University of Coimbra, Coimbra, Portugal  
joana@fe.uc.pt

<sup>2</sup> CeBER, Faculty of Economics, University of Coimbra, Coimbra, Portugal  
{joana, pgodinho}@fe.uc.pt, pedro.torres@uc.pt

**Abstract.** Predicting in advance whether a given customer will end his relationship with a company has an undeniable added value for all organizations, since targeted campaigns can be prepared to promote customer retention. In this work, six different methods using machine learning have been investigated on the retail banking customer churn prediction problem, considering predictions up to 6 months in advance. Different approaches are tested and compared using real data. Out of sample results are very good, even with very challenging out-of-sample sets composed only of churners, that truly test the ability to predict when a customer will churn. The best results are obtained by stochastic boosting, and the most important variables for predicting churn in a 1–2 months horizon are the total value of bank products held in recent months and the existence of debit or credit cards in another bank. For a 3–4 months horizon, the number of transactions in recent months and the existence of a mortgage loan outside the bank are the most important variables.

**Keywords:** Churn · Retail banking · Machine learning · Validation

## 1 Introduction

Customer churn prediction is an important component of customer relationship management since it is less profitable to attract new customers than to prevent customers to abandon the company [1, 2]. Companies that establish long term relationships with their customers can focus on customers' requirements, which is more profitable than looking for new customers [3]. This may also enable companies to reduce service costs [2] and enhances opportunities to cross and up sale [1]. Long-term customers are also less likely to be influenced by competitors' marketing campaigns [4], and tend to buy more and to spread positive word-of-mouth [2]. Customers that abandon the company may influence others to do the same [5]. Actually, customers that churn because of "social contagion" may be harder to retain [6]. Thus, building predictive models that enable the identification of customers showing high propensity to abandon are of crucial importance, since they can provide guidance for designing campaigns aiming to persuade customers to stay [7].

Machine learning (ML) techniques have been used for churn prediction in several domains. For an overview of the literature after 2011 see [1, 7]. Few publications

consider churn prediction in the financial sector or retail banking. In the work presented in [8], only 6 papers considered the financial sector. In another literature analysis focusing on applications of business intelligence in banking, the authors conclude that credit banking is the main application trend [9]. Words like “retention” or “customer relationship management” are not the most relevant term frequencies found.

Most of the referenced work consider applications in the telecommunication sector (see, for instance, [10–15]). Applications of churn prediction methods in other sectors can also be found. Coussement et al. [16] consider churn prediction in a newspaper publishing company. Miguéis et al. [17] consider the retailing context. Buckinx and Van den Poel [18] predict partial defection in fast-moving consumer goods retailing. Predicting churn in the logistics industry is considered in [19].

As aforementioned, research focusing on retail banking customer churn prediction has been scarce. The prediction of churn in this context has some differentiating features. There is a strong competitive market and a single customer can have several different retail banking service providers at the same time, which offer more or less the same type of products with similar costs. The relationship between the customer and the service provider is, in general, of the non-contractual type, meaning that the customer has control over the future duration and type of the relationship with the company. The customer has control over the services he wants to acquire. A customer can, most of the time, leave the bank without informing it of this intention, making it harder to distinguish between churners or simply inactive customers.

In a non-contractual setting, it is often difficult to know what is the proper churn definition that should be considered and different definitions can have different economic impacts for the companies, namely considering the cost and impact of marketing campaigns against churn [20]. In the retail banking sector, the situation is even more complicated, since the cost of terminating the relationship with a service provider can be very different for different types of customers: it is negligible for many, but not for all. Customers that have signed loans for home purchase, for instance, can incur in significant costs for terminating their relationship with the bank, and this relationship is much more similar to a contractual type one than for other types of customers.

Churn prediction in the banking sector has been addressed not only by machine learning approaches but also by survival analysis models. Mavri and Ioannou [21] use a proportional hazard model to determine the risk of churn behavior, considering data from Greek banking. Their objective is not to predict whether a specific customer will churn or not, but rather to understand what features influence the switching behavior of bank customers. Survival analysis is also used by Larivière and Van den Poel [22], who consider Kaplan-Meier estimates to determine the timing of churn. The authors use data from a large Belgian financial service provider. The authors also study the influence of product cross-selling on the customer propensity to churn. The same authors apply machine learning approaches, namely random forests, to the same dataset and compare its performance with logistic regression for customer retention and profitability prediction. They conclude that the best results are achieved by random forests [23].

Glady et al. [24] propose the use of customer lifetime value as a measure for classifying churners in the context of a retail financial service company. A churner is defined as a customer with a decreasing customer lifetime value. The authors develop a new loss function for misclassification of customers, based in the customer lifetime

value. The authors apply decision trees, neural networks and logistic regression models as classifiers. The predictor variables used are of the type RFM (recency, frequency and monetary). Xie et al. [25] apply random forests to churn prediction considering banks in China, integrating sampling techniques and cost-sensitive learning, penalizing more the misclassification of the minority class and achieving very high accuracy rates. De Bock and Van den Poel [26] develop a model able to predict partial churn for a European bank (the closing of a checking account by a customer, not taking into consideration whether that customer has other accounts or services in the bank) in the next 12 months. The authors conclude that the most important features influencing churn prediction are related to RFM variables. Verbeke et al. [7] consider the importance of the comprehensibility and justifiability of churn prediction models, and use AntMiner+ and Active Learning Based Approach for Support Vector Machine rule extraction. The concern with interpretability is also shared by [26], that use an ensemble classifier based upon generalized additive models and apply it to a set of six datasets. The model interpretability is illustrated by considering a case study with a European bank. De Bock and Poel [27] apply two different forest-based models, Rotation Forest and RotBoost, to four of these six real-life datasets. Gür Ali and Arıtürk [8] describe a dynamic churn prediction framework in the context of private banking including environmental variables in addition to customers' behavior attributes. Churn prediction in a defined future time window is tackled using binary classifiers, and the performance is compared with Cox Regression. Shirazi and Mohammadin [28] focus on the customer retiree segment of the Canadian retail banking market. They develop a model that considers not only data directly related with the customers' interactions with the bank but also with the customers' behavior (tracking of customers' behavior on various websites, for instance).

Churn prediction for bank credit card customers is addressed by several authors. Farquad et al. [29] apply support vector machine in conjunction with Naïve Bayes Trees for rule generation. This hybrid approach outperforms other approaches in the computational tests performed using a Latin American bank dataset. Lin et al. [30] use rough set theory to extract rules that explain customer churn. Nie et al. [31] apply logistic regression and decision trees to a dataset from a Chinese bank, reaching the conclusion that logistic regression slightly outperforms decision trees.

In this work, six machine learning techniques are investigated and compared to predict churn considering real data from a retail bank. Individual results obtained by each methodology are also compared with the results obtained by applying survival analysis tools. The objective is to develop a methodological framework capable of predicting not only which customers will churn but also when they will churn, considering a future horizon of six months. The models aim at predicting which will be the customers churning in the next month, two months from now, and so on.

By employing different machine learning tools and evaluating the models using a large retail banking dataset, this study makes several contributions to the literature:

- It is focused on retail banking, an industry in which research has been scarce, and shows how retail banking data can be leveraged by structuring data in new ways.
- It defines of a global framework for churn prediction that not only identifies the customers that are more likely to churn, but also predicts when they are going to

churn, allowing a continuous reassessment of the churn probabilities for different time horizons. It can therefore be a useful tool for supporting the management of resources and scheduling of campaigns in retail banks.

- It proposes a new validation methodology that truly assesses the prediction accuracy, in a highly-biased dataset. This methodology considers the use of balanced out-of-sample sets that are only composed of churners, with different churning times.
- It presents computational results using real data, showing that it is possible to achieve high levels of predictive accuracy. This means that the model is useful in a real-life context.

Although a large dataset is used, the time horizon of the sample dataset (two years) is an important constraint. This limitation was overcome by creating different records based on different rolling time windows, but a dataset covering a longer time period might provide more insights.

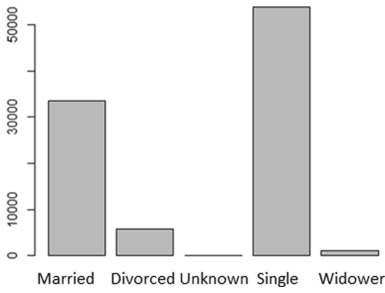
This paper is organized as follows: Sect. 2 describes the methodology applied. Section 3 presents the main computational results. Section 4 presents a discussion of the results and some ideas for future research.

## 2 Exploratory Data Analysis

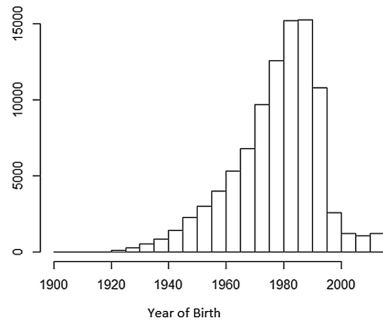
The data used in this study considers information related to more than 130 000 customers of a retail bank, including all the monthly customer interactions with the bank for two years: product acquisition, product balances, use of bank services, in a total of 63 attributes for each month. All the data is anonymized, preventing the identification of customers. Personal data considers age, location, marital status, employment situation, date of bank account opening and the way in which the customer opened the account (using the bank online platform, in a bank branch or any other way). Considering the large number of attributes characterizing the customers' interactions with the bank, some of these attributes were consolidated. The final set of attributes considered were: total value of bank products held by the customer, total value of personal loans, total value of mortgage loans, number of insurance policies held, total number of transactions of any kind (debit or credit cards, bank transfers, deposits, etc.), binary values stating whether the customer had a mortgage loan or personal loan in another bank, and whether he has debit or credit cards in another bank.

Regarding the sociodemographic characterization of customers, most of them are single, followed by those that are married, as shown in Fig. 1. Most customers were born between 1980 and 1990 (Fig. 2).

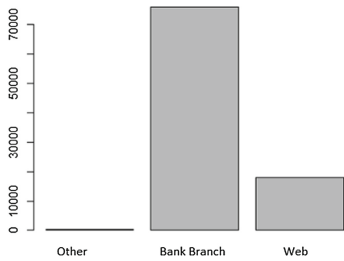
Most of the customers opened an account using a bank branch (Fig. 3), but the oldest ones have mostly opened their accounts using the web platform (Fig. 4).



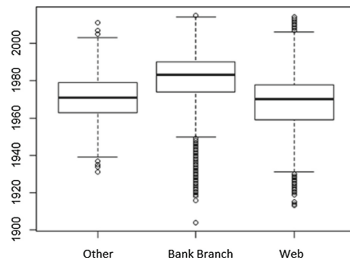
**Fig. 1.** Distribution of customers by marital status



**Fig. 2.** Distribution of customers by year of birth



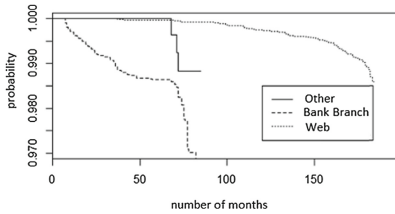
**Fig. 3.** Distribution of customers by acquisition channel



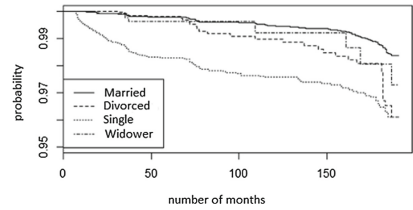
**Fig. 4.** Distribution of customers by acquisition channel and year of birth

Less than 1% of the customers have churned during the time horizon corresponding to the available data. In order to better understand what characterizes a churner, a survival analysis was performed, using Cox Regression. The acquisition channel, marital status, age, level of academic education and number of descendants showed to be statistically significant attributes ( $p < 0.01$ ) in explaining churn. Survival curves were plotted considering different customer related attributes. Figures 5 and 6 illustrate two of those survival plots. It was possible to conclude that customers that begin their relationship with the bank using the web are less prone to churn. Customers that are single present a higher probability of churning in earlier stages of their involvement with the bank. Older customers are less prone to churn. Regarding the academic education, less educated customers are the ones that churn earlier. These results are fully aligned with other similar results found in the literature. The results in [32], for instance, show that older people are more likely to stay with the bank, and that more educated people are more likely to be loyal.

All the attributes were pre-processed in the following way: all quantitative attributes were centred and scaled and the natural logarithm was applied, due to the skewness that the histograms of most attributes presented.



**Fig. 5.** Survival curves considering acquisition channel



**Fig. 6.** Survival curve considering marital status

### 3 Prediction Models and Methodology

The bank has its own definition of churn: it wants to be able to predict customers that fall below a certain threshold in terms of the relationship with the bank: a customer is a churner if he does not interact in any way with the bank for 6 consecutive months, the balance of assets in the bank is smaller than or equal to 25 € and the balance of debts is also smaller than or equal to 25 €. It is considered that the time of churn is the first time these conditions are simultaneously met.

The problem of predicting up to six months in advance the customers that will churn is treated as a classification problem with two classes (churners and non-churners). The dataset is highly unbalanced since the minority class (churners) constitutes less than 1% of the total dataset.

The objective is to predict which customers will churn and when they will churn, considering a future time horizon of 6 months. Therefore, six different models are built. One model will predict the customers that will churn in the next month. Another model will predict which customers will churn two months from now, and so on. These predictions will be based on the data corresponding to the past six months, along with the known customer attributes.

To train the machine learning models, different datasets considering rolling time windows dependent on the prediction horizon were created. For example, consider that the model will make predictions for the next month. As there are 24 months of data available, the first time window will consider the initial seven months of data: the first six months will originate the values of the explanatory variables (attributes). The information of whether a customer has churned or not in the seventh month will be the output variable (to be predicted). Then this time window is rolled one time period forward. Next, all records with explanatory variables considering months 2 to 7 and the output variable calculated for month 8 are added to the dataset. This is performed until there are no more data available. In the case of one month ahead prediction, the last time window includes months 18–24. Building the datasets in this way means that a given customer will contribute more than one record for the global dataset. Actually, a customer that will churn at a given month will contribute with records classified as “non-churner” until the churn actually occurs, and one record classified as “churner” corresponding to the month in which he churns.

Figures 7 and 8 illustrate the use of these rolling time windows for predictions made one and two months ahead, respectively.

Due to the highly-unbalanced feature of the data, it is important to balance the dataset before feeding the data into the machine learning models. An undersampling strategy was chosen. All the “churner” samples were considered and a subset of the “non-churner” samples, composed by a number of samples that matched the “churner” ones, was randomly selected. Burez and Van den Poel [33] study the problem of class imbalance that is often present in churn prediction datasets, since the churners are usually the minority class. The authors state that undersampling can be useful, improving prediction accuracy, and they find no advantages in using advanced sampling techniques. Other authors also state that undersampling seems to present advantages when compared with oversampling [34, 35].

To use the churners data as most as possible, leave-one-out cross validation was chosen to test the models and to calculate accuracy metrics. From a given data set, all data related to a given customer is removed. The model is trained with the remaining data, and then it tries to predict what happens with the removed customer. For each model, average accuracy (number of correct predictions divided by the total number of predictions) and average area under the curve (AUC, defined as the area below a plot of the true positive rate against the false positive rate for the different thresholds) are calculated. In a balanced dataset, a perfect forecast will produce a 100% value for the accuracy and for the AUC, whereas a completely random prediction will produce a value of about 50% for each of these metrics.

When predicting churn, one possible criticism is that customers that churn and that do not churn have different features that facilitate the classification process. To test even further the machine learning models applied, the creation of the datasets has been slightly changed in a way that emphasizes the ability of the methods to accurately predict the time of churn. As, for churners, several records will usually be available (one record classified as “churner”, for the month when churn occurred, and all other records classified as “non-churners”), sample sets created exclusively with customers that have churned during the 24 months considered are created. The methods will have to find out if they have churned in the next month, next two months, and so on.

Six different methods were used, using the indicated R packages: random forests (RF-randomForest), support vector machine (SVM-kernlab), stochastic boosting (SB-ADA), logistic regression (LR-Rpart), classification and regression trees (CART-Rpart), multivariate adaptive regression splines (MARS-Earth).

Months											
1	2	3	4	5	6	<u>7</u>	8	9	10	...	24
1	2	3	4	5	6	7	<u>8</u>	9	10	...	24
1	2	3	4	5	6	7	8	<u>9</u>	10	...	24

**Fig. 7.** Rolling window for building the dataset considering predictions one month in advance (values from the explanatory variables will come from the months in grey background, output variables will be calculated using data from the underlined months).

Months											
1	2	3	4	5	6	7	<u>8</u>	9	10	...	24
1	2	3	4	5	6	7	8	<u>9</u>	10	...	24
1	2	3	4	5	6	7	8	9	<u>10</u>	...	24

**Fig. 8.** Rolling window for building the dataset considering predictions two months in advance (values from the explanatory variables will come from the months in grey background, output variables will be calculated using data from the underlined months).

RF are a combination of tree predictors [36], each one built considering a random subset of the available data, and contributing with one vote to the final result, calculated by majority voting. RF with 1000 trees were created. SVM are non-probabilistic supervised classifiers that find the hyperplane such that the nearest training data points belonging to different classes are as distant as possible. Radial basis gaussian functions were used. SB considers the application of simple classifiers, each one using a different version of the training data with different weights associated [37]. CART does a recursive partitioning for classification, building a binary tree by deciding, for each node, what is the best splitting variable to choose (minimizing a classification error). MARS creates a piecewise linear model using the product of spline basis functions [38].

The methodological framework used is described in pseudo-code:

1.  $n \leftarrow$  forecasting horizon
2. Build the dataset for this forecasting horizon:
  - a. Initialize the rolling time window with  $t \leftarrow 1$ .  $Dataset \leftarrow \{ \}$
  - b. For each customer, build a sample considering the explanatory variables retrieved from data from month  $t$  to  $t + 5$ , and determine the value of the output variable (churner/non-churner) by looking at month  $t + 5 + n$ . Include samples from all available customers in the *Dataset*.
  - c. If  $t + 5 + n < 24$  then  $t \leftarrow t + 1$  and go to 2.b. Else go to 3.
3. Repeat for 5 times:
  - a. Create a set with all the samples that correspond to churners.
  - b. Include in this set an identical number of samples randomly retrieved from the “non-churner” records.
  - c. For each existing customer  $i$  in this set:
    - (1) Remove customer  $i$  from the set.
    - (2) Train the machine learning method with the remaining samples.
    - (3) For each record belonging to customer  $i$  predict the corresponding output value.

Repeating five times the third step is important for two reasons: on one hand, some of the machine learning methods that are going to be applied have a random behaviour, so it is important to assess average behaviour instead of reaching conclusions based on a single run; on the other hand, as the sampling set is built using a random sampling procedure, it is also important to see if the results are sensitive to this sampling.



## 4 Computational Results

All methods were assessed considering classification accuracy and AUC, for each forecast horizon, and considering the two situations: datasets built considering churners and non-churners, and datasets built considering churners only.

These methods were also compared with the use of Cox Regression and the corresponding survival functions.

Survival analysis is able to predict whether a customer will churn or not during the considered time horizon (24 months). Considering a threshold of 0.5, so that if the survival probability is less than this threshold it predicts churn, survival analysis has an average accuracy of 80.73% and an average AUC of 89% (considering balanced sets). However, it is not capable of accurately predicting when a customer will churn. The average error is 6 months.

Table 1 presents the accuracy obtained when the dataset considers both churners and non-churners. A threshold of 0.5 was considered for all methods. Average values and standard deviations are shown for all the six methods tested, and for all the forecasting horizons. These values are related with out-of-sample testing only. As expected, with the increase of the forecasting horizon the precision deteriorates. The method that presents the best results in general is SB. The values of the standard deviation show that the methodology is not very sensitive to the random components of the procedures that were employed.

Table 2 presents the results for AUC. These results are also very good for SB. The methods are being tested considering balanced datasets, which means that a random classifier would have an AUC near 50%.

When the datasets were exclusively composed of churners that did churn sometime during the 24 months period, the quality of the results deteriorates, as expected. However, the accuracy and AUC are still very good (Table 3 and Table 4). SB continues to be the method presenting the best results for AUC. CART is slightly better than SB when it comes to accuracy. These results show that it is not only possible to accurately classify churners and non-churners, but it is also possible to accurately predict when they are going to churn.

It is possible to better understand the attributes that define whether a customer will churn or not by looking at the importance of each variable in the predicted outcome. For SB, the importance of each variable is related with the number of times it is selected for boosting.

Considering SB models predicting churn one and two months ahead, the most important variables are the total value of bank products held by the customer in the past 3 months, along with the existence of debit or credit cards in another bank. Considering models that predict churn 3 to 4 months ahead, the preceding attributes continue to be important, but the most important ones are the number of transactions in the past 2 months and the information of whether the customer has a mortgage loan outside the bank. These variables are mostly the same considering the two different types of sample sets (with churners and non-churners and churners only).

**Table 1.** Accuracy considering a dataset with churners and non-churners (grey cells- the best value for the corresponding forecasting horizon; italic and bold values – the worst results for the corresponding forecasting horizon).

	RF		SVM		SB		LR	
Future	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
<b>1 month</b>	88,39%	0,59%	<b><i>82,39%</i></b>	0,42%	89,07%	0,67%	84,50%	0,72%
<b>2 months</b>	89,37%	0,41%	<b><i>80,39%</i></b>	0,98%	88,46%	1,23%	82,28%	0,65%
<b>3 months</b>	87,88%	0,87%	77,85%	1,08%	88,33%	0,87%	<b><i>76,53%</i></b>	5,54%
<b>4 months</b>	87,67%	1,31%	78,50%	0,84%	88,06%	0,72%	<b><i>75,99%</i></b>	5,95%
<b>5 months</b>	86,15%	1,25%	78,12%	0,53%	87,43%	0,67%	<b><i>75,94%</i></b>	5,55%
<b>6 months</b>	85,68%	0,95%	<b><i>75,55%</i></b>	2,18%	86,05%	0,76%	78,74%	1,44%
	CART		MARS					
Future	Average	Standard Deviation	Average	Standard Deviation				
<b>1 month</b>	86,55%	2,12%	85,52%	0,91%				
<b>2 months</b>	85,59%	2,46%	85,55%	1,12%				
<b>3 months</b>	87,88%	0,89%	85,84%	1,37%				
<b>4 months</b>	85,60%	1,96%	86,11%	1,49%				
<b>5 months</b>	85,32%	2,98%	83,80%	2,47%				
<b>6 months</b>	85,83%	0,65%	82,96%	1,84%				

**Table 2.** AUC considering a dataset with churners and non-churners (grey cells- the best value for the corresponding forecasting horizon; italic and bold values – the worst results for the corresponding forecasting horizon).

	RF		SVM		SB		LR	
Future	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
<b>1 month</b>	96,19%	0,27%	<b><i>88,32%</i></b>	0,99%	96,53%	0,31%	91,90%	0,60%
<b>2 months</b>	96,29%	0,17%	<b><i>86,80%</i></b>	0,55%	96,14%	0,39%	89,92%	0,46%
<b>3 months</b>	95,52%	0,35%	84,99%	0,44%	95,78%	0,34%	<b><i>81,32%</i></b>	9,51%
<b>4 months</b>	95,38%	0,61%	85,41%	1,19%	95,70%	0,44%	<b><i>80,74%</i></b>	9,55%
<b>5 months</b>	94,03%	0,48%	85,62%	1,53%	95,20%	0,52%	<b><i>80,51%</i></b>	9,43%
<b>6 months</b>	93,38%	0,41%	<b><i>82,82%</i></b>	1,52%	94,45%	0,40%	86,10%	1,26%
	CART		MARS					
Future	Average	Standard Deviation	Average	Standard Deviation				
<b>1 month</b>	90,59%	1,18%	94,79%	0,36%				
<b>2 months</b>	90,06%	1,82%	94,70%	0,68%				
<b>3 months</b>	85,25%	4,76%	94,71%	0,69%				
<b>4 months</b>	81,87%	7,91%	94,36%	0,76%				
<b>5 months</b>	86,25%	6,72%	93,65%	1,20%				
<b>6 months</b>	87,15%	1,97%	92,35%	1,21%				

**Table 3.** Accuracy considering a dataset with churners only (grey cells- the best value for the corresponding forecasting horizon; italic and bold values – the worst results for the corresponding forecasting horizon).

	RF		SVM		SB		LR	
Future	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
<b>1 month</b>	82,97%	0,53%	70,46%	3,14%	82,48%	0,56%	<i>67,36%</i>	2,46%
<b>2 months</b>	79,62%	0,91%	70,53%	3,90%	<i>81,19%</i>	0,76%	<b>64,42%</b>	5,09%
<b>3 months</b>	78,27%	1,44%	63,82%	5,29%	79,85%	1,15%	<b>63,62%</b>	3,62%
<b>4 months</b>	77,25%	0,90%	<b>60,35%</b>	5,44%	78,13%	0,92%	62,23%	10,02%
<b>5 months</b>	76,56%	1,40%	<b>60,67%</b>	4,26%	76,63%	1,13%	62,61%	2,38%
<b>6 months</b>	73,19%	1,48%	<b>56,71%</b>	6,05%	75,61%	0,91%	57,93%	2,71%
	CART		MARS					
Future	Average	Standard Deviation	Average	Standard Deviation				
<b>1 month</b>	78,26%	3,90%	80,30%	1,59%				
<b>2 months</b>	79,31%	1,31%	80,23%	1,34%				
<b>3 months</b>	81,13%	1,06%	79,97%	1,74%				
<b>4 months</b>	80,48%	0,88%	78,02%	1,17%				
<b>5 months</b>	78,70%	1,22%	77,27%	3,06%				
<b>6 months</b>	78,66%	0,88%	75,39%	3,25%				

**Table 4.** AUC considering a dataset with churners only (grey cells- the best value for the corresponding forecasting horizon; italic and bold values – the worst results for the corresponding forecasting horizon).

	RF		SVM		SB		LR	
Future	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
<b>1 month</b>	89,73%	0,31%	78,24%	3,14%	90,49%	0,70%	72,79%	1,87%
<b>2 months</b>	87,37%	0,72%	77,02%	3,90%	89,25%	0,58%	<b>69,59%</b>	6,98%
<b>3 months</b>	85,93%	1,21%	71,75%	5,29%	88,31%	0,71%	<b>68,82%</b>	3,78%
<b>4 months</b>	84,69%	0,26%	68,14%	5,44%	86,84%	0,52%	<b>65,25%</b>	12,29%
<b>5 months</b>	84,02%	0,48%	<b>66,06%</b>	4,26%	85,85%	0,72%	67,06%	2,50%
<b>6 months</b>	79,38%	1,04%	<b>58,48%</b>	6,05%	84,00%	0,82%	59,88%	2,48%
	CART		MARS					
Future	Average	Standard Deviation	Average	Standard Deviation				
<b>1 month</b>	78,76%	3,90%	87,62%	0,62%				
<b>2 months</b>	80,31%	1,31%	87,61%	0,41%				
<b>3 months</b>	78,90%	1,06%	86,79%	1,31%				
<b>4 months</b>	78,19%	0,88%	85,05%	1,14%				
<b>5 months</b>	77,82%	1,22%	84,50%	0,83%				
<b>6 months</b>	75,71%	0,88%	82,10%	1,37%				

## 5 Conclusions

In this study, a methodology is investigated that allows a retail bank to produce each month a list of customers that are more likely to churn in the next six months, detailing which ones are likely to churn in each month ahead. This tool can be an important asset for determining focused retaining campaigns that will only be targeting customers with a high probability of leaving, and that the bank wants to keep. There is also the possibility that some of these customers are not interesting to the bank, so churning should not be avoided in these cases. The importance of not only correctly identifying churners but also deciding which of them to include in retaining campaigns, with the objective of maximizing profit, is addressed in [39], for instance.

The classification threshold considered was 0.5, so that no bias was introduced in the computational results shown. However, this value can be optimized to account for different costs associated with incorrectly predicting a churner or a non-churner. Predicting that a given customer will churn, when in fact he is not going to, can imply the cost of an unnecessary retaining action. Predicting that a customer is not a churner when indeed he is will imply the cost of losing future profits associated with this customer, which can be more significant to the bank. This analysis should be made to decide on the best threshold value to consider.

In the computational tests, SB had the best overall performance. Attributes related to the relationship that the customer has with other bank institutions are important for churn prediction.

In future work, an integrated methodology between churner prediction and the prediction of the next product to buy can be developed. This would help a company understanding what kind of product should offer to a possible churner in order to retain him as a customer. Future studies could also use other evaluation metrics that include profits [1].

This study only superficially addresses the motivations to churn. Computational results show that it is somehow related to the relationship of each customer with other banks. Inspired by Óskarsdóttir et al. [10], it would be interesting to better understand the motivations to churn, including “social contagion”.

**Acknowledgments.** This study has been funded by national funds, through FCT, Portuguese Science Foundation, under project UIDB/00308/2020.

## References

1. 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. *Eur. J. Oper. Res.* **269**, 760–772 (2018)
2. Ganesh, J., Arnold, M.J., Reynolds, K.E.: Understanding the customer base of service providers: an examination of the differences between switchers and stayers. *J. Mark.* **64**, 65–87 (2000)
3. Reinartz, W.J., Kumar, V.: The impact of customer relationship characteristics on profitable lifetime duration. *J. Mark.* **67**, 77–99 (2003)

4. Colgate, M., Stewart, K., Kinsella, R.: Customer defection: a study of the student market in Ireland. *Int. J. Bank Mark.* **14**, 23–29 (1996)
5. Nitzan, I., Libai, B.: Social effects on customer retention. *J. Mark.* **75**, 24–38 (2011)
6. Verbraken, T., Bravo, C., Weber, R., Baesens, B.: Development and application of consumer credit scoring models using profit-based classification measures. *Eur. J. Oper. Res.* **238**, 505–513 (2014)
7. Verbeke, W., Martens, D., Mues, C., Baesens, B.: Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Syst. Appl.* **38**, 2354–2364 (2011)
8. Gür Ali, Ö., Antürk, U.: Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Syst. Appl.* **41**, 7889–7903 (2014)
9. Moro, S., Cortez, P., Rita, P.: Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Syst. Appl.* **42**, 1314–1324 (2015)
10. Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., Vanthienen, J.: Time series for early churn detection: Using similarity based classification for dynamic networks. *Expert Syst. Appl.* **106**, 55–65 (2018)
11. Tsai, C.-F., Chen, M.-Y.: Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Syst. Appl.* **37**, 2006–2015 (2010)
12. Tsai, C.-F., Lu, Y.-H.: Customer churn prediction by hybrid neural networks. *Expert Syst. Appl.* **36**, 12547–12553 (2009)
13. Huang, B., Kechadi, M.T., Buckley, B.: Customer churn prediction in telecommunications. *Expert Syst. Appl.* **39**, 1414–1425 (2012)
14. 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. *Comput. Electr. Eng.* **38**, 1808–1819 (2012)
15. Vafeiadi, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, KCh.: A comparison of machine learning techniques for customer churn prediction. *Simul. Model. Pract. Theor.* **55**, 1–9 (2015)
16. Coussement, K., Benoit, D.F., Van den Poel, D.: Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Syst. Appl.* **37**, 2132–2143 (2010)
17. Miguéis, V.L., Camanho, A., Falcão e Cunha, J.: Customer attrition in retailing: an application of multivariate adaptive regression splines. *Expert Syst. Appl.* **40**, 6225–6232 (2013)
18. Buckinx, W., Van den Poel, D.: Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *Eur. J. Oper. Res.* **164**, 252–268 (2005)
19. Chen, Kuanchin., Hu, Ya-Han, Hsieh, Yi-Cheng: Predicting customer churn from valuable B2B customers in the logistics industry: a case study. *Inf. Syst. e-Bus. Manage.* **13**(3), 475–494 (2014). <https://doi.org/10.1007/s10257-014-0264-1>
20. Clemente-Ciscar, M., San Matías, S., Giner-Bosch, V.: A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings. *Eur. J. Oper. Res.* **239**, 276–285 (2014)
21. Mavri, M., Ioannou, G.: Customer switching behaviour in Greek banking services using survival analysis. *Manag. Financ.* **34**, 186–197 (2008)
22. Larivière, B., Van den Poel, D.: Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: the case of financial services. *Expert Syst. Appl.* **27**, 277–285 (2004)

23. Larivière, B., Van den Poel, D.: Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst. Appl.* **29**, 472–484 (2005)
24. Glady, N., Baesens, B., Croux, C.: Modeling churn using customer lifetime value. *Eur. J. Oper. Res.* **197**, 402–411 (2009)
25. Xie, Y., Li, X., Ngai, E.W.T., Ying, W.: Customer churn prediction using improved balanced random forests. *Expert Syst. Appl.* **36**, 5445–5449 (2009)
26. De Bock, K.W., Van den Poel, D.: Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Syst. Appl.* **39**, 6816–6826 (2012)
27. De Bock, K.W., den Poel, D.V.: An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Syst. Appl.* **38**, 12293–12301 (2011)
28. Shirazi, F., Mohammadi, M.: A big data analytics model for customer churn prediction in the retiree segment. *Int. J. Inf. Manage.* **48**, 238–253 (2018)
29. Farquad, M.A.H., Ravi, V., Raju, S.B.: Churn prediction using comprehensible support vector machine: an analytical CRM application. *Appl. Soft Comput.* **19**, 31–40 (2014)
30. Lin, C.-S., Tzeng, G.-H., Chin, Y.-C.: Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Syst. Appl.* **38**, 8–15 (2011)
31. Nie, G., Rowe, W., Zhang, L., Tian, Y., Shi, Y.: Credit card churn forecasting by logistic regression and decision tree. *Expert Syst. Appl.* **38**, 15273–15285 (2011)
32. Van den Poel, D., Larivière, B.: Customer attrition analysis for financial services using proportional hazard models. *Eur. J. Oper. Res.* **157**, 196–217 (2004). [https://doi.org/10.1016/S0377-2217\(03\)00069-9](https://doi.org/10.1016/S0377-2217(03)00069-9)
33. Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. *Expert Syst. Appl.* **36**, 4626–4636 (2009)
34. Chen, C.: *Using Random Forest to Learn Imbalanced Data*. University of California, Berkeley (2004)
35. Weiss, G.M.: Mining with rarity: a unifying framework. *ACM SIGKDD Explor. Newslett.* **6**, 7 (2004)
36. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
37. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. *Ann. Stat.* **28**, 337–407 (2000)
38. Friedman, J.H.: Multivariate adaptive regression splines. *Ann. Stat.* 1–67 (1991)
39. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B.: New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *Eur. J. Oper. Res.* **218**, 211–229 (2012)