Health club customer dropout membership*

Pedro Sobreiro, Javier Berrocal, Domingos Martinho, José Garcia Alonso

Extremadura University

21 julho, 2022

Abstract

Customer retention is fundamental to improve the organizations profits. Existing approaches normaly explore static models predicting when the customer will dropout. In this paper we propose a different approach considering that the customer dropout risk changes over time using clusters to improve the model performance. We explore a survival model using random forests with and with out clusters in a dataset of 5209 customers. The model using clusters improved the performance significantly. This paper shows that the use of clusters should be considered to identify dropout patterns to support the timming when the dropout occurs considering the cluster where the customer is. Improving the performance as less errors and is expected to contribute to the identification when should be developed retention strategies.

*Corresponding address: sobreiro@esdrm.ipsantarem.pt. Quality of Life Research Centre, Polytechnic Institute of Santarém, Portugal

1 Introduction

Customer retention is a problem is a problem being addressed using the dropout prediction as an insight to identify customers that could dropout. The customers database is the most valuable asset that the organizations possess (Athanassopoulos 2000). The problem of retention is not new, in a seminal paper Copeland (1923) address the problem of brand loyalty and in marketing research (Mellens, Dekimpe, and Steenkamp 1996).

The advantage of developing some strategies in retention are supported in the idea that the costs of customer retention are lower than customer acquisition (Edward and Sahadev 2011; Fornell and Wernerfelt 1987). The increase in the profits with the reduction of 5% of the dropout could represent almost a duplication of the profits (Reichheld 1996).

The development of a customer retention strategy could be supported in the identification of the customers that will dropout (Alboukaey, Joukhadar, and Ghneim 2020a). However, dropout has two underlying scenarios contractual and non-contractual settings {Gupta et al. (2006); Ascarza (2018)}, in a contractual business the customer needs to renew their contracts to continue its usage (Ascarza and Hardie 2013), against non contractual where the firm has to infer if the customer is still active However, in contractual settings the customer dropout represents an explicit ending of a relationship which is more penalizing than non contractual settings (Risselada, Verhoef, and Bijmolt 2010). This has implications to the profitability of the organizations increasing marketing costs and reducing sales (Amin et al. 2017).

The anticipation of the dropout allows the development of countermeasures to reduce customer churn. Several studies address the problem related to customer retention trying to improve the profitability (Coussement and Van den Poel 2009; Devriendt, Berrevoets, and Verbeke 2019; García, Nebot, and Vellido 2017). Existing organizations are addressing this problem by shifting their target from capturing new customers to preserving existing ones (García et al. 2017), considering that investments in retention strategies are more profitable than acquiring new customers (Coussement and Van den Poel 2009).

The approaches normally employed use a dependent variable representing dropout or non-dropout, without considering a dynamic perspective that the dropout risk changes over time (Alboukaey, Joukhadar, and Ghneim 2020b). The survival models try to solve this limitation (Routh, Roy, and Meyer 2020) capturing a temporal dimension of the customer dropout (Perianez et al. 2016). Perianez et al. (2016) used survival analysis to predict also when the dropout will occur.

Other studies proposed also the integration of several algorithms to improve the performance in the prediction of the dropout such the usage of clusters combined with churn prediction (Gök, Özyer, and Jida 2015; Hung, Yen, and Wang 2006; Vijaya and Sivasankar 2019). The approach relies in the assumption that combining the customers in different clusters allows the improvement of the prediction accuracy. Vijaya and Sivasankar (2019) suggested the adoption hybrid models combining more than one classier are achieving increased performance compared to those using single classifiers.

There are several challenges around the timing related to dropout, or considering the dynamic behavior of the customer in the intent to drop out (Alboukaey et al. 2020a). The importance of understanding when dropout will occur and the risk when discarding the temporal perspective of the problem seems to be an element that should be addressed. Few studies considered this (Burez and Vandenpoel 2008; Perianez et al. 2016). This shows an opportunity to address the importance of the timeframe and its influence on the efficiency of the model and also evalute if the combination of clusters could improve the performance.

Survival analysis, which origin as stands in biomedical statistics, its especially well-suited to studying the timing of events in longitudinal data (Singer and Willett 1993). Survival analysis is a class of statistical methods modelling the occurrence and timing of an event, such as the customer dropout. Survival analysis allow us to examine not only if an event occurred but also how long it took to occur. A primary value of survival analysis, however, is to compare dropout probability for individuals classified with theoretically relevant variables. The survival methods have enjoyed and increasing popularity in several disciplines ranging from medicine to economics (Singer and Willett 1993).

Random Survival Forests does not make the proportional hazards assumption (Ehrlinger 2016) and has the flexibility to model survivor curves that are of dissimilar shapes for contrasting groups of subjects. Random Survival Forest is an extension of Random Forest allowing efficient non-parametric analysis of time to event data (Breiman 2001). This characteristics allow us to surpass the Cox Regression limitation of the proportional hazard assumption, requiring to exclude variables which not fulfill the model assumption. It was shown by Breiman (2001) that ensemble learning can be further improved by injecting randomization into the base learning process - a method called Random Forests.

Some studies employed also a combination of clusters with the churn prediction (Gök et al. 2015; Hung et al. 2006; Vijaya and Sivasankar 2019), where the customers where grouped in clusters to improve the accuracy within each cluster. Clusters are approaches using unsupervised algorithms to group elements with similar characteristics. This unsupervised method has been widely used employing approaches such Hierarchical Clustering (Saunders 1980), K-Means (Vijaya and Sivasankar 2019), or Random Forest Clustering (Breiman 2001). Jafari-Marandi et al. (2020) explored an approach combining clustering methods in parallel to a classification.

In this study, we investigate if an hybrid approach using clusters and random survival forests which have never been used in to predict membership in a health club using existing data improves the prediction accuracy. Our paper is organized as follows. In the next section, we address our research methodology, using survival analysis to identify the survival probability along the time, how we address the problem of determining the clusters and the performance of the model to predict customer survival within each identified cluster.

2 Survival analysis

Survival analysis focus in the analysis of the time until an event of interest, and exploring its relationship with different factors. The main advantage is related to the concept of censoring, indicating that observations that are not complete related to the event of interest, e.g. customers that didn't dropout yet, which are incorporated in the analysis. This means that there are customers still active for which we don't know if the event of dropout has occurred, which is called censorship. Survival models take censoring into account and incorporate this uncertainty, instead of predicting the time of event such in regression models, the survival models allow to predict the probability of an event happens at a particular time.

The time of dropout is represented by T, which is a non-negative random variable, indicating the time period of the event occurring for a randomly selected individual from the population, representing the probability of an event to occur each time period given that has not already occurred in a previous time period, known as discrete-time hazard function (Singer and Willett 1993). The survival function represents the probability of an individual surviving after time t, S(t) = P(T $>$ t), t $\geq$ 0, with the properties S(0) = 1,S(∞) = 0. The distribution function is represented with F, defined as F(t) = P(T $\leq$ 0), for t $\geq$ 0. The function of probability density represented with f where:

$$f(t) = \lim_{dt \to 0} \frac{P[t \leq T < t + dt]}{dt} \tag{1}$$

$f(t)dt$ represents the probability of an event occurring in the moment t. The need to represent the distribution evolution of the death probability along the time, uses to the hazard function, represented as:

$$\lambda(t) = \lim_{dt \to 0} \frac{P[t \leq T < t + dt | T \geq t]}{dt} \tag{2}$$

The determination of the survival curves is based in the following elements: (1) the total value of observations removed during the time period (e.g. days, months or years), either by dropout or by censorship; (2) observations that composed the sample of the study; (3) customers who had not yet dropped out at any given time. The survival probability until the time period ii (p_i) is calculated with:

$$p_i = \frac{r_i - d_i}{r_i} \tag{3}$$

Where r_i is the number of individuals that survived at the beginning of the period, d_i the number of individuals who left during the period. The survival time estimate was also taken considering the month in which it is found (estimated). Cox's allow test difference between survival times. The advantage in using survival analysis was that allow us to detect if the risk of an event differs systematically across different people, using specific predictors. The coefficients in a Cox regression were related to the hazard, where a positive value represents a worse prognosis and the opposite, negative value a better prognosis. The advantage of survival analysis was that allow us to include information of covariates that were censored up to the censoring event.

The Cox PH model assumes the covariates to be time independent, for example gender and age when where retrieved do not change over time (Schober and Vetter 2018) Because the Cox model requires the hazards in both groups to be proportional, researchers are often asked to "test" whether hazards are proportional (Stensrud and Hernán 2020). Considering this we explored other approach that allow us to develop this analysis without the proportional hazard assumptions, the survival trees.

3 Survival Trees

Survival trees are methods based in tree based models based on Random Forest (Breiman 2001). Random survival forests is an ensemble method for analysis of right-censored data (Ishwaran et al. 2008), using randomization to improve the performance. Random survival forests follows this framework (Ishwaran et al. 2008):

1. Draw B random samples of the same size from the original dataset with replacement. The samples that are not drawn are said to be out-of-bag (OOB). Grow a survival tree on each of the b = 1 , . . . , B samples.
2. At each node, select a random subset of predictor variables and find the best predictor and splitting value that provide two subsets (the daughter nodes) which maximizes the difference in the objective function.
3. Repeat step 2 recursively on each daughter node until a stopping criterion is met.
4. Calculate a cumulative hazard function (CHF) for each tree and average over all CHFs for the B trees to obtain the ensemble CHF.
5. Compute the prediction error for the ensemble CHF using only the OOB data.

In each node is selected a predictor x from a random selected predicted variables and split value c (one unique value of x). Each sample i if assigned the daughter right node if $x_i \leq c$ or left if $x_i \geq c$, then is calculated the logrank such as:

$$L(x,c) = \frac{\sum_{i=1}^{N} \left(d_{i,1} - Y_{i,1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i,1}}{Y_i} \left(1 - \frac{Y_{i,1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}} \tag{4}$$

Where:

- j: Daughter node, $j \in \{1, 2\}$
- $d_{i,j}$: Number of events at time t_i in daughter node j
- $Y_{i,j}$: Number of elements that had the event or are in risk at time t_i in daughter node j
- d_i: Number of events at time t_i, such $d_i = \sum_j d_{i,j}$
- Y_i: Number of elements that experienced an event or are at risk at t_i so $Y_i = \sum_j Y_{i,j}$

We loop every x and c until find x^* that satisfy $|L(x^*, c^*)| \geq |L(x, c)|$ for every x and c. The model performance was determined with the concordance probability (C-index), Brier Score (BS) and Mean Absolute Error (MAE) (Wang, Li, and Reddy 2017). The feature importance was determined calculating the difference between the true class label and noised data (Breiman 2001).

The BS is used to evaluate the predicted accuracy of the survival function at a given time t. Representing the average square distance between the survival status and the predicted survival probability, where the value 0 is the best possible outcome.

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\left(0 - \hat{S}(t, \vec{x}_i)\right)^2 \cdot \mathbb{1}_{T_i \leq t, \delta_i = 1}}{\hat{G}(T_i^-)} + \frac{\left(1 - \hat{S}(t, \vec{x}_i)\right)^2 \cdot \mathbb{1}_{T_i > t}}{\hat{G}(t)} \right) \tag{5}$$

The model should have a Brier score below 0.25. Considering that if $\forall i \in \llbracket 1, N \rrbracket, \hat{S}(t, \vec{x}_i) = 0.5$, then $BS(t) = 0.25$.

4 Methodology

To simplify the analysis, the survival probabilities are presented as a survival curve. The survival curve is a representation of the survival probabilities corresponding to a time where the events are observed (Bland and Altman 1998). The survival analysis was conducted using the package Lifelines (Davidson-Pilon 2021). Where dropout is a binary value where one represent churn and zero not churn. The dropout happens when a member does not have a payment.

The random survival forest was developed using the package PySurvival (Fotso et al. 2019). PySurvival is an open source python package for Survival Analysis modeling - the modeling concept used to analyze or predict when an event is likely to happen. The model was built with with 70% of the data for training and 30% for testing.

The model performance was determined with the concordance probability (C-index), Brier Score (BS) and Mean Absolute Error (MAE) (Wang et al. 2017). The feature importance was determined calculating the difference between the true class label and noised data (Breiman 2001).

The BS measure the average discrepancies between the status (dropout/non-dropout) and the estimated probabilities at a given time. The Integrated Brier Score (IBS) was used to calculate the performance in all available times (from t_1 to t_{max}) as:

$$IBS = \int_{t_1}^{t_{\max}} BS^c(t)dw(t) \tag{6}$$

Representing the average square distance between the survival status and the predicted survival probability, where the value 0 is the best possible outcome.

The calculation of he number of clusters used the package mclust (Scrucca et al. 2016) using the Bayesian Information Criterion (BIC). The model that gives the minimum BIC score can be selected as the best model (Schwarz 1978) simplifying the problem related to choosing the number of components and identifying the structure of the covariance matrix, based on modelling with multivariate normal distributions for each component that forms the data set (Akogul and Erisoglu 2016).

The hybrid approach was develop as follows:

1. Identify the optimal number of clusters using (Scrucca et al. 2016)
2. Fit the model using the identified number of clusters
3. Estimate for each element the cluster
4. for each cluster follow the framework proposed by Ishwaran et al. (2008) to calculate the random survival model

5 Dataset

In this study, data from 5,209 fitness customers was analysed (mean age = 27.88, SD=11.80 years) from a Portuguese fitness centre. The data was collected from software e@sport (Cedis, Portugal) between 2014 and 2017. The information retrieved was: Age of the participants in years; Sex (0-female, 1-male); Non-attendance days before dropout; Total amount billed; Average number

Table 1: Summary statistics of features used

Characteristic	**N = 5,209**
Age in years, Mean (SD)	28 (12)
Male or female, %	35%
dayswfreq, Mean (SD)	76 (102)
tbilled, Mean (SD)	155 (155)
maccess, Mean (SD)	0.89 (0.76)
freeuse, %	4.9%
nentries, Mean (SD)	29 (41)
cfreq, %	
2	1.3%
4	2.4%
6	0.2%
7	96%
months, Mean (SD)	9 (8)
dropout, %	88%

of visits per week; Total number of visits; Weekly contracted accesses; Number of registration renewals; Number of customer referrals; Registration month; Customer enrolment duration; and status (dropout/non-dropout). Dropout event occur when customer communicate the intention to terminate the contract or did not pay the monthly fee during 60 days.

Table 1 shows data's summary statistics. The average age is 27.9 ± 11.8, the entries are 29 ± 41.2 with an inscription period of 9 ± 8.2 months.

Figure 1 shows the distribution of the dropout considering the number of months of membership.

6 Results

The table 2 depicts the data of the survival time of the customers during the first months, the results showed that the customers have a survival probability of 24.44% at 12 months (column p_i - likelihood probability) with a median survival time of 10 months (column estimated_survival). The survival probability at 6 months was 54.5%, representing an risk of dropout of 45.5% with a estimated survival of 6 months.

Figure 2 shows the Kaplan Meier survival curve customers considering the number of months of membership (x axis) and survival probability (y axis). The customer dropout is very high in the first 12 months, ranging from a survival probability of 54% after the first 6 months until 24% after 12 months.

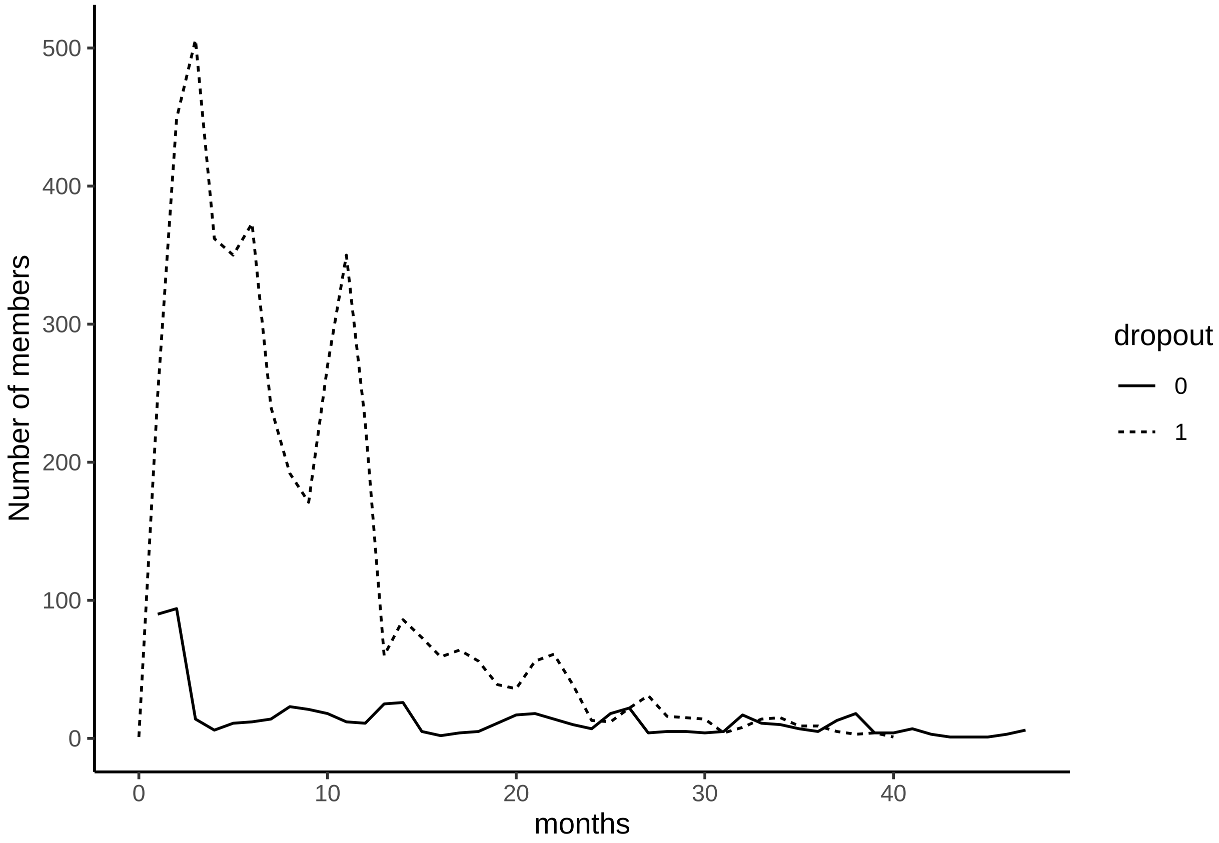


Figure 1: Number of members by month

Table 2: Determination of the survival time probabilities

event_at	removed	observed	censored	entrance	at_risk	estimated_survival	prob
0	1	1	0	5209	5209	7	1.000
1	339	249	90	0	5208	7	0.952
2	543	449	94	0	4869	7	0.864
3	520	506	14	0	4326	7	0.763
4	368	362	6	0	3806	7	0.691
5	361	350	11	0	3438	6	0.620
6	385	373	12	0	3077	6	0.545
7	254	240	14	0	2692	5	0.496
8	215	192	23	0	2438	6	0.457
9	192	171	21	0	2223	6	0.422
10	288	270	18	0	2031	6	0.366
11	362	350	12	0	1743	9	0.293
12	240	229	11	0	1381	10	0.244
13	85	60	25	0	1141	9	0.231
14	112	86	26	0	1056	9	0.212
15	78	73	5	0	944	9	0.196
16	61	59	2	0	866	10	0.183
17	68	64	4	0	805	10	0.168
18	61	56	5	0	737	9	0.155
19	50	39	11	0	676	9	0.146
20	53	36	17	0	626	9	0.138
21	74	56	18	0	573	10	0.124
22	75	61	14	0	499	11	0.109
23	49	39	10	0	424	11	0.099
24	20	13	7	0	375	10	0.096

Note:

Removed – the sum of customers with dropout and that are censored; Censored – the event did not occur during the period of this data, collection; Risk of Dropout – number of customers at risk of, dropout; pi – survival probability; Estimated Survival - months to survive in the sports facility.

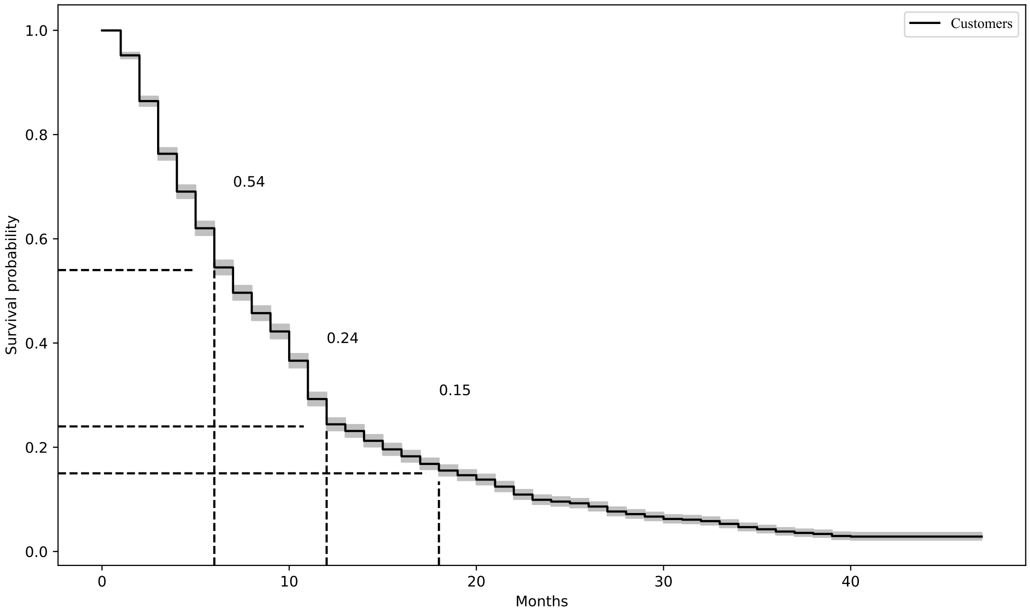


Figure 2: Survival probabilities

Figure 3 shows the survival by gender. The survival curves by gender are very similar, both types of customers present a behavior that is not very different.

Figure 4 shows the survival by contracted frequency. Customers with contracted frequency of 6 and 4 times a week have higher survival probabilities, against lower of customers with contracted frequencies of 7 and 2 times a week. Survival curves allow to explore tendencies related to survival to extract actionable knowledge.

The proportional hazard assumptions failed in the following variables: age p<0.01, cfreq p<0.01, dayswfreq p<0.01, tbilled p<0.01, freeuse p<0.01, nentries p<0.01.

6.1 Survival trees

To evaluate the performance of the random survival forest we have calculated the concordance probability (C-index), IBS and Mean Absolute Error (MAE). The IBS presents an accuracy along the 12 months of 0.08 (figure 5).

The actual versus predicted model presents the actual and predicted customers which dropout during the 40 months, which as an average absolute error of 7.5 customers (figure 6).

Table 3 shows features importance calculated according (Breiman 2001), where the percent increase

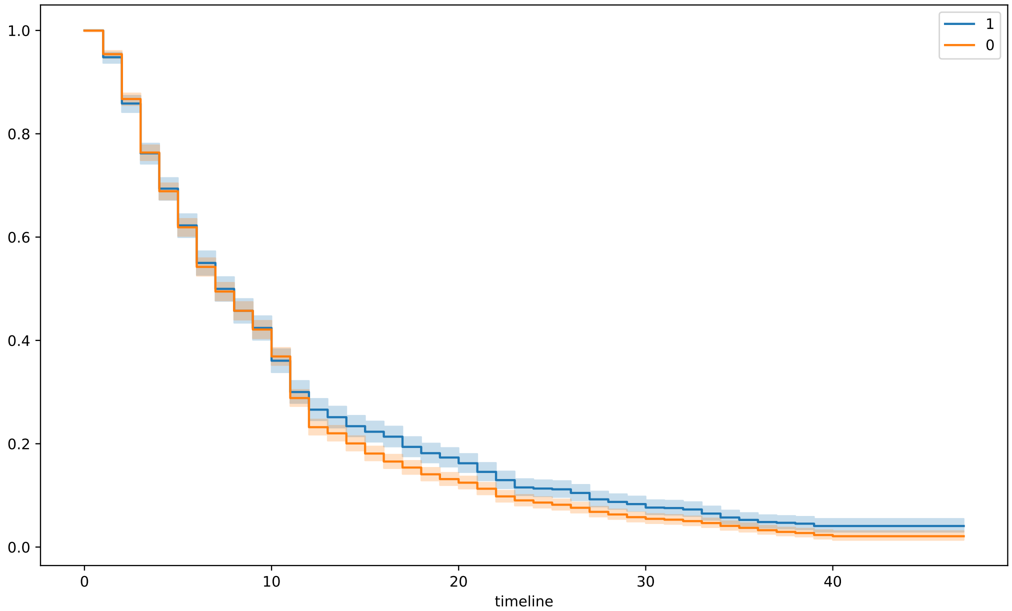


Figure 3: Survival by gender

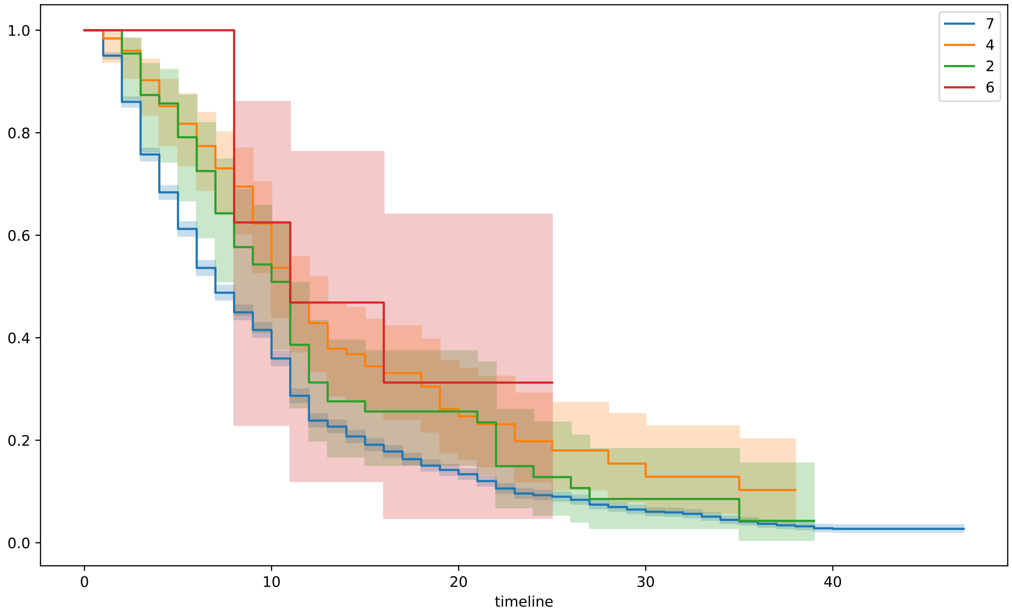


Figure 4: Survival by contracted frequency

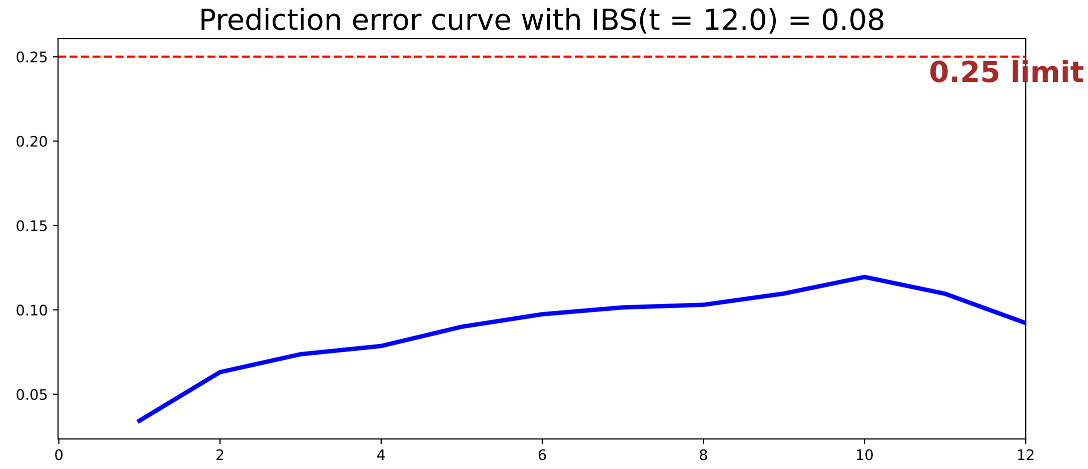


Figure 5: Model performance

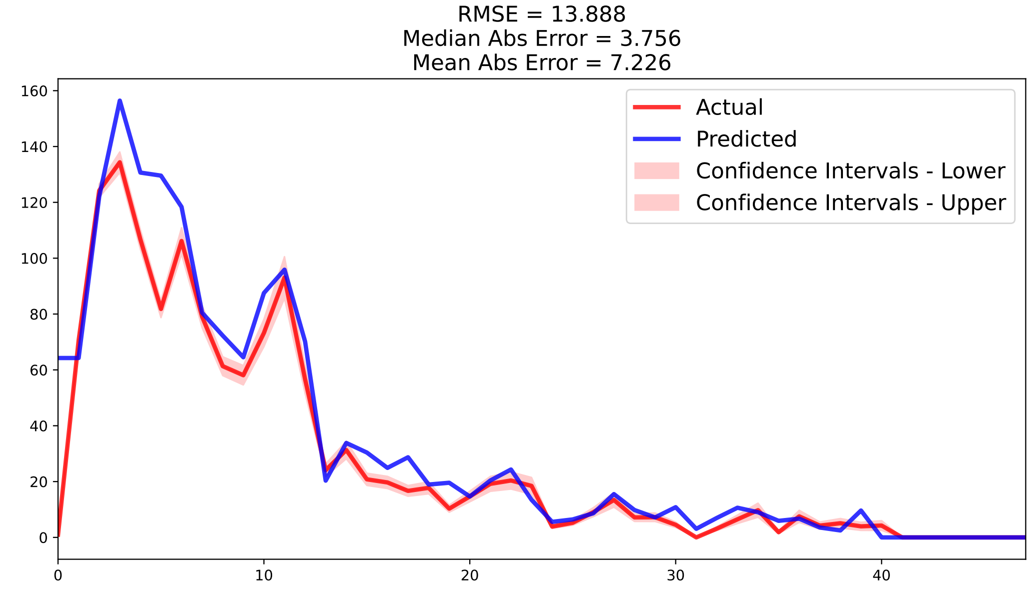


Figure 6: Conditional survival forest

Table 3: Features importance in the survival model

feature	importance	pct_importance
tbilled	7.9434176	0.2845915
dayswfreq	5.3046172	0.1900503
nentries	4.7644466	0.1706974
maccess	4.6711845	0.1673561
freeuse	3.2174794	0.1152737
cfreq	2.0105041	0.0720310
age	-0.2880629	0.0000000
sex_1	-1.5439344	0.0000000

in misclassification rate as compared to the out-of-bag rate (with all variables intact), out-of-bag is a bootstrap aggregating (subsampling with replacement to create training samples for the model to learn from) where two independent sets are created. One set, the bootstrap sample, data chosen to be "in-the-bag" by sampling with replacement and the out-of-bag is all data not chosen in the sampling process. The most important variable is the $dayswfreq$, followed by $tbilled$ and $nentries$, compared with the $cfreq$, age, and sex.

The prediction is very similar to the actual value. The model accuracy is very high with a root mean square error of 14. The mean absolute error mean was 7.23 customers, and the median absolute error was 3.76.

6.2 Survival trees based model with clusters

In this approach we have created clusters and applied the survival trees within each cluster. The determination of the clusters using the BIC criterion where the EEV model: 7 clusters -57159.24; 6 clusters -63937.59; and 4 clusters -77088.81 figure 7 shows the determination of the number of clusters using BIC, also the elbow analysis available in figure 8. An optimal number of clusters was considered of five. Considering that was the value after the average distortion was flattened.

The calculation of the clusters to each member in the dataset was developed considering, 7,6 and 4 clusters. However after the executions KMeans only performed with 3 clusters.

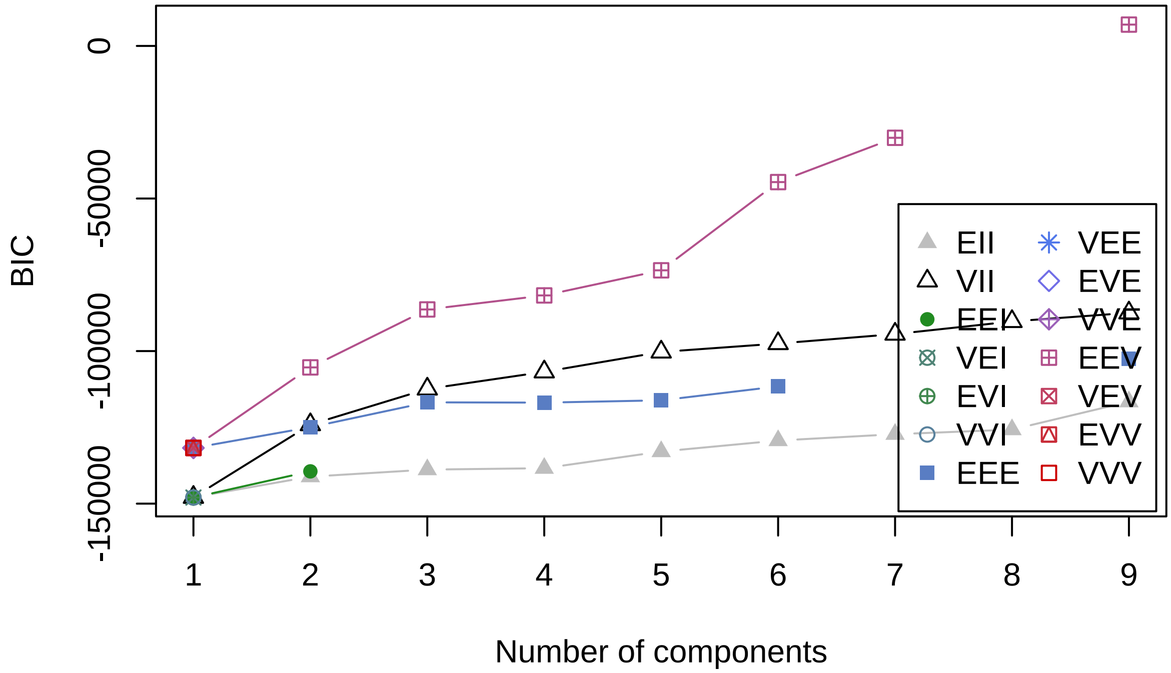


Figure 7: Analysis number of clusters

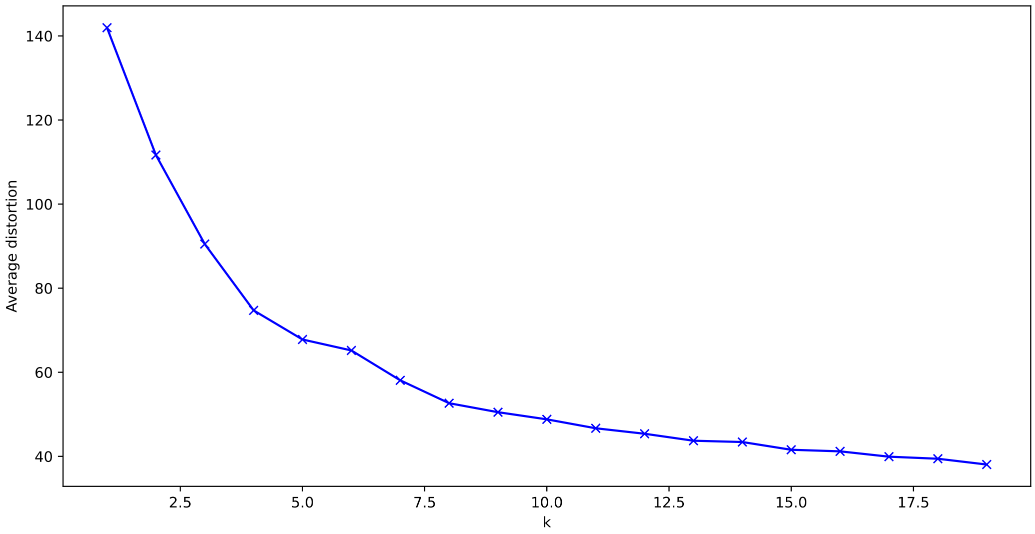
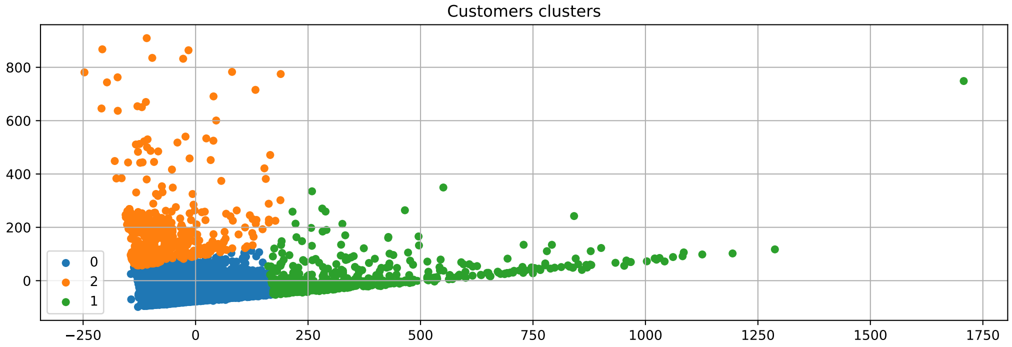
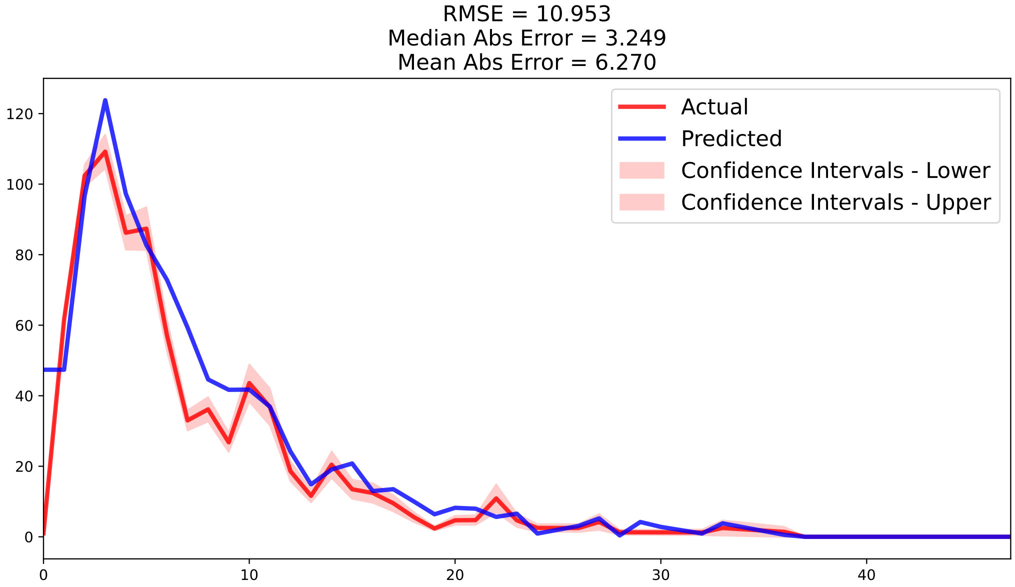
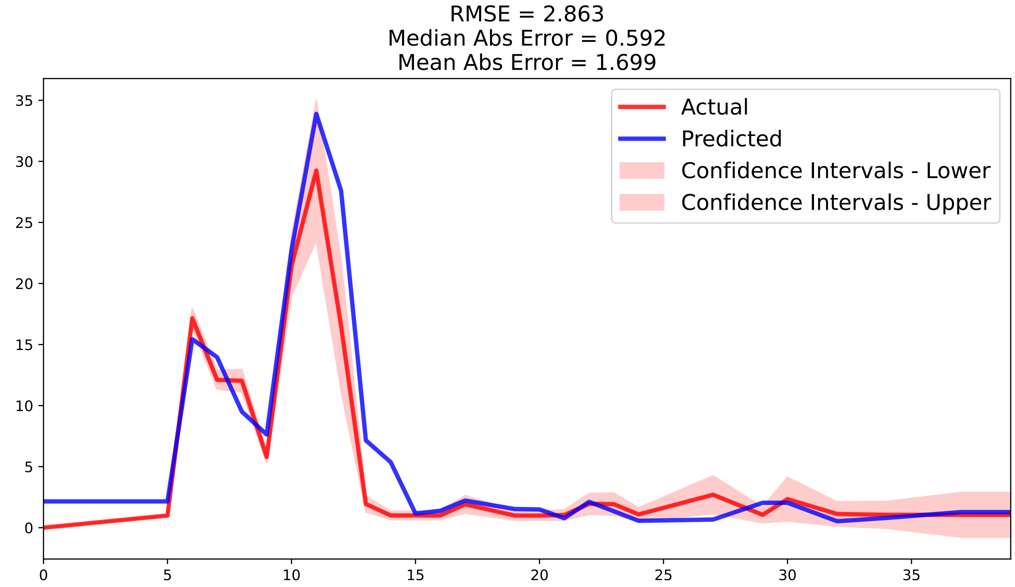


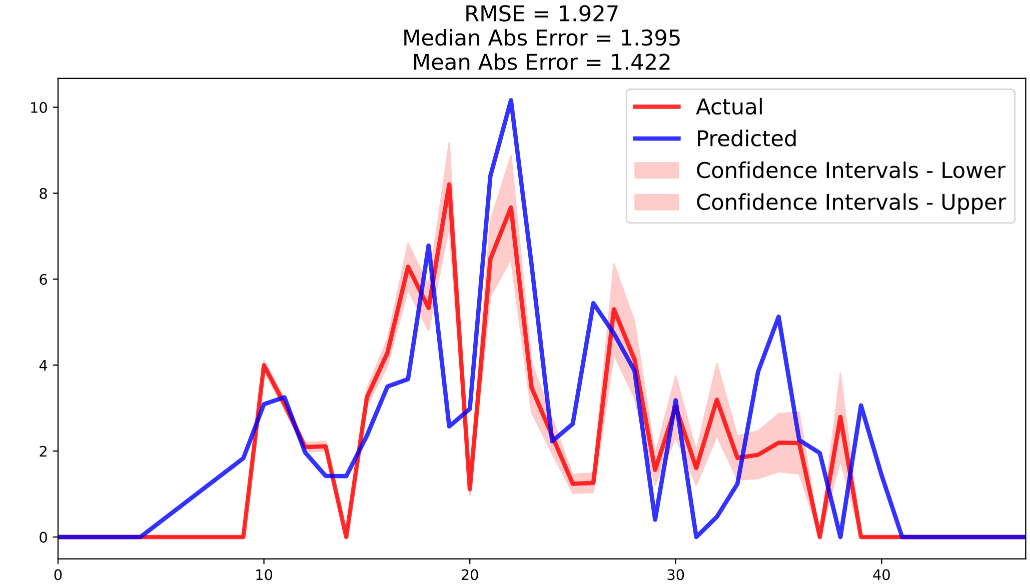
Figure 8: Elbow analysis



The model accuracy is very high in the first years. The prediction is very similar to the actual value. The absolute error mean of 6 customers.


RMSE = 10.953
Median Abs Error = 3.249
Mean Abs Error = 6.270
Actual
Predicted
Confidence Intervals - Lower
Confidence Intervals - Upper
120
100
80
60
40
20
0
0
10
20
30
40


RMSE = 2.863
Median Abs Error = 0.592
Mean Abs Error = 1.699
Actual
Predicted
Confidence Intervals - Lower
Confidence Intervals - Upper
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35



The performance of the cluster 1 the IBS presents an accuracy of 0.06 (figure 9) along all time. The actual versus predicted model presents the actual and predicted customers which dropout during the 40 months, which as an mean absolute error of 1.5 customers, the mean median absolute error was 0.61 and the Root Mean Square Error of 2.8 (figure 10).

The features importance in the survival model cluster 1 (table 4) identify the three most relevant features to predict survival $maccess$, $tbilled$, and $dayswfreq$. The features with lower relevance were $freeuse$, sex and $cfreq$.

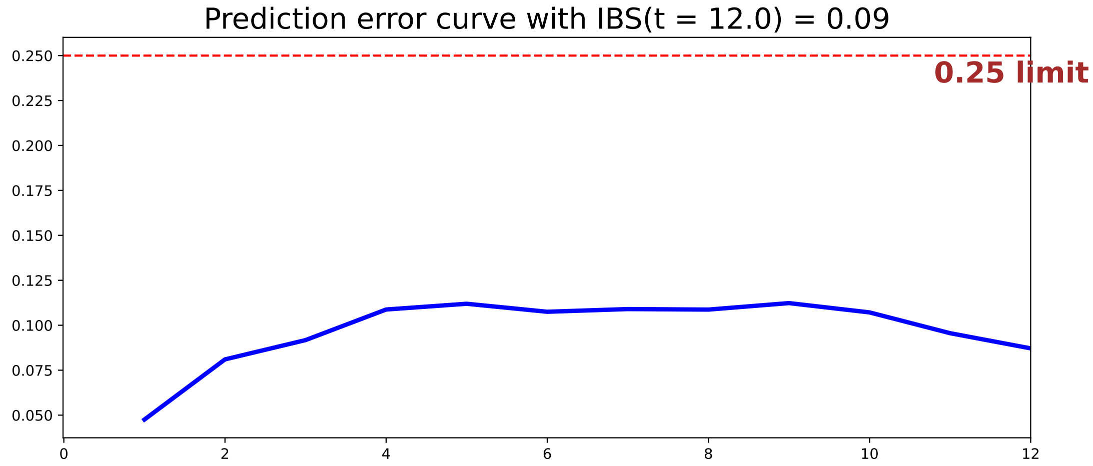


Figure 9: Model performance cluster 1

The performance of the cluster 2 the IBS presents an accuracy along time 0.09 (figure 11) along all

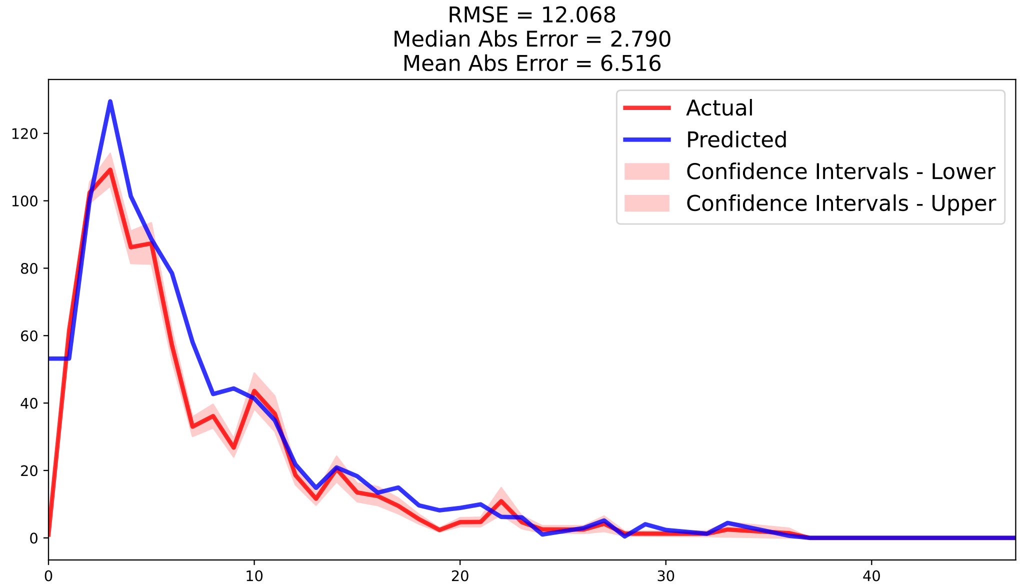


Figure 10: Conditional survival forest cluster 1

Table 4: Features importance in the survival model with cluster 1

feature	importance	pct_importance
tbilled	6.7827606	0.2644082
freeuse	5.7521637	0.2242331
maccess	4.4870321	0.1749152
dayswfreq	3.9224018	0.1529046
nentries	3.0360807	0.1183537
cfreq	0.8262370	0.0322087
age	0.4607052	0.0179594
sex_1	0.3852319	0.0150173

time. The actual versus predicted model presents the actual and predicted customers which dropout during the 40 months, which as an mean absolute error of 6.5 customers, the mean median absolute error was 2.3 and the Root Mean Square Error of 11.79 (figure 12). The features importance in the survival model cluster 2 (table 5) identify the three most relevant features to predict survival $daysw freq$, $tbilled$, and $freeuse$. The least relevant were $nentries$, $cfreq$, and age.

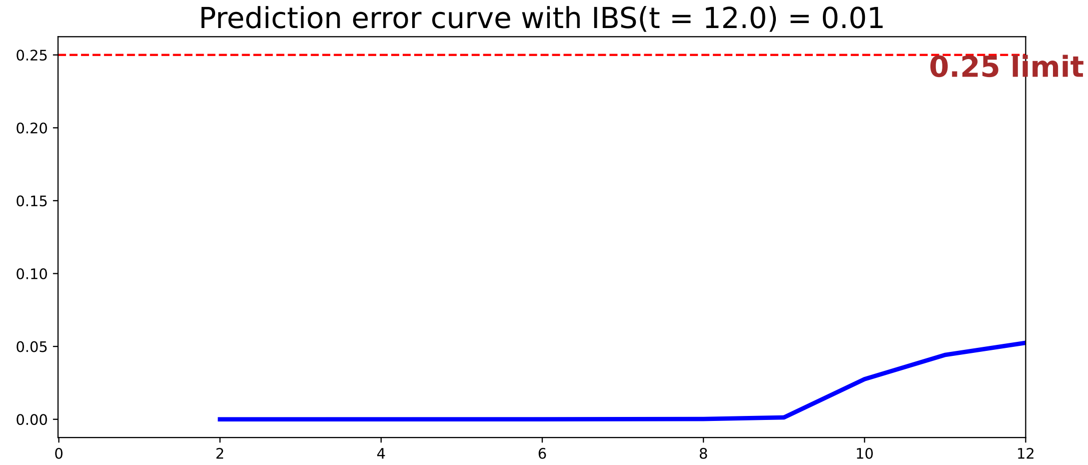


Figure 11: Model performance cluster 2

The performance of the cluster 3 the IBS presents an accuracy along time 0.01 (figure 13) along all time. The actual versus predicted model presents the actual and predicted customers which dropout during the 40 months, which as an mean absolute error of 1.5 customers, the mean median absolute error was 1.2 and the Root Mean Square Error of 2.08 (figure 14). The features importance in the survival model cluster 3 (table 6) identify the three most relevant features to predict survival $tbilled$, $daysw freq$, and $nentries$. The least relevant were sex, age, and $cfreq$.

The features importance in the survival model (table 6) identify the three most relevant features to predict survival $daysw freq$, $tbilled$, and $nentries$.

6.2.1 Model Comparison

Table 7 shows the performance of both approaches, with or without clusters. The RMSE in the clusters 1, 2 and 3 is lower than not using clusters to predict dropout. Overall the performance improved. The performance is also better using mean and median.

The model accuracy without clusters is very high with a root mean square error of 13, the mean absolute error mean was 7.53 customers, and the median absolute error was 4.04. The model using

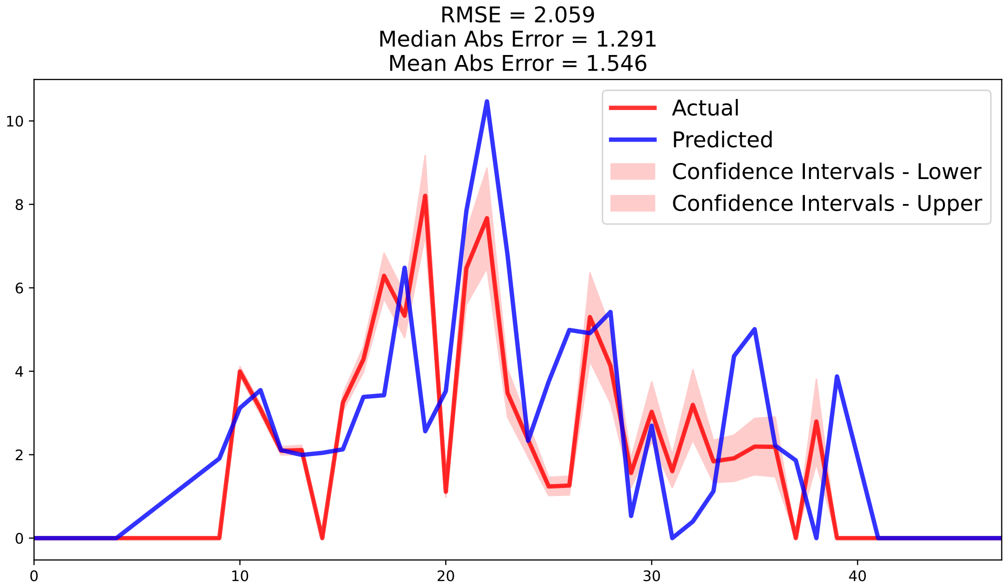


Figure 12: Conditional survival forest cluster 2

Table 5: Features importance in the survival model with cluster 2

feature	importance	pct_importance
tbilled	6.6524419	0.2970936
maccess	4.0288064	0.1799238
dayswfreq	3.9304155	0.1755297
nentries	2.7923452	0.1247043
freeuse	2.4090838	0.1075881
age	1.4997470	0.0669777
sex_1	1.0789000	0.0481829
cfreq	-0.6615769	0.0000000

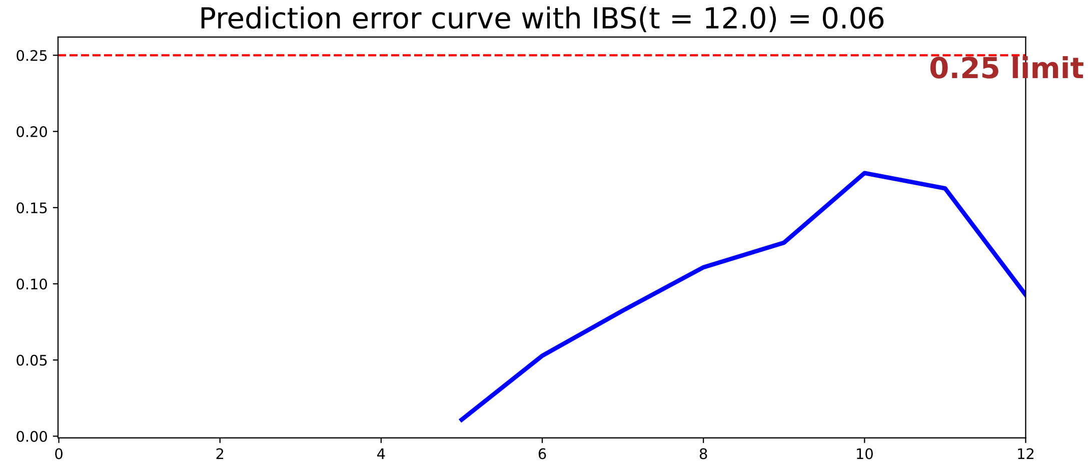


Figure 13: Model performance cluster 3

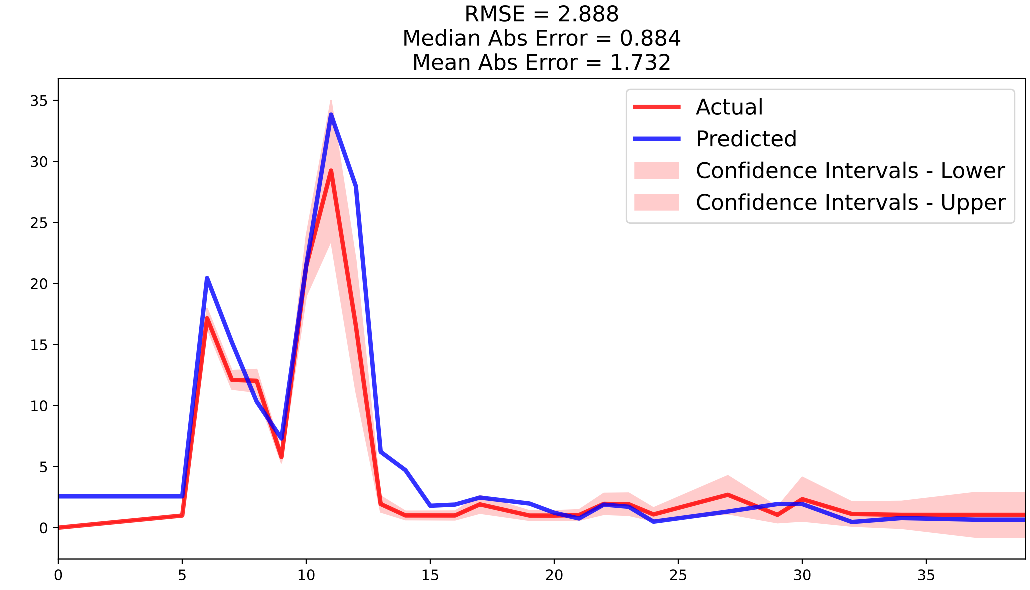


Figure 14: Conditional survival forest cluster 3

Table 6: Features importance in the survival model with cluster 3

feature	importance	pct_importance
tbilled	3.7920500	0.3033014
nentries	2.8590522	0.2286769
dayswfreq	2.3304981	0.1864014
maccess	1.8131070	0.1450186
age	1.5533976	0.1242461
sex_1	0.1544767	0.0123556
freeuse	0.0000000	0.0000000
cfreq	0.0000000	0.0000000

Table 7: Performance of prediction in each cluster

cluster	rmse	mean	median
0	10.9534010098597	3.24949079763695	6.27043826620657
2	2.86277805553651	0.592331070836667	1.69865648584295
1	1.92696670365884	1.39451757989365	1.42197114146461
w/cluster	13.8883596176347	3.75609199994054	7.22583527582812

clusters had an mean absolute error of 1.5 customers, the mean median absolute error was 1.2 and the Root Mean Square Error of 2.08. The performance using clusters improved significantly.

7 Conclusion

This paper investigated the customer dropout in a Health Club organization, using a dynamic perspective that the dropout risk varies along the time. Exploring two approaches, using a survival model based on random forests with or without clusters. The model using clusters allowed to combine the customers in different clusters, an hybrid approach. Based on this performance the proposed model using clusters allows to improve the accuracy on the survival model allowing to target approaches considering the timing when the dropout occurs, considering the clusters where the customer is. Is very important for managers use this information to improve their retention strategies.

References

Akogul, Serkan and Murat Erisoglu. 2016. “A Comparison of Information Criteria in Clustering Based on Mixture of Multivariate Normal Distributions.” *Mathematical and Computational*

Applications 21(3):34.

Alboukaey, Nadia, Ammar Joukhadar, and Nada Ghneim. 2020b. “Dynamic Behavior Based Churn Prediction in Mobile Telecom.” *Expert Systems with Applications* 162:113779.

Alboukaey, Nadia, Ammar Joukhadar, and Nada Ghneim. 2020a. “Dynamic Behavior Based Churn Prediction in Mobile Telecom.” *Expert Systems with Applications* 162:113779.

Amin, Adnan, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Khalid Alawfi, Amir Hussain, and Kaizhu Huang. 2017. “Customer Churn Prediction in the Telecommunication Sector Using a Rough Set Approach.” *Neurocomputing* 237:242–54.

Ascarza, Eva. 2018. “Retention Futility: Targeting High-Risk Customers Might Be Ineffective.” *Journal of Marketing Research* 55(1):80–98.

Ascarza, Eva and Bruce G. S. Hardie. 2013. “A Joint Model of Usage and Churn in Contractual Settings.” *Marketing Science* 32(4):570–90.

Athanassopoulos, Antreas D. 2000. “Customer Satisfaction Cues to Support Market Segmentation and Explain Switching Behavior.” *Journal of Business Research* 47(3):191–207.

Bland, J. M. and D. G. Altman. 1998. “Survival Probabilities (the Kaplan-Meier Method).” *BMJ (Clinical Research Ed.)* 317(7172):1572.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45(1):5–32.

Burez, J. and D. Vandenpoel. 2008. “Separating Financial from Commercial Customer Churn: A Modeling Step Towards Resolving the Conflict Between the Sales and Credit Department.” *Expert Systems with Applications* 35(1-2):497–514.

Copeland, Melvin T. 1923. “Relation of Consumers’ Buying Habits to Marketing Methods.” *Harvard Business Review* 1(3):282–89.

Coussement, Kristof and Dirk Van den Poel. 2009. “Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers.” *Expert Systems with Applications* 36(3, Part 2):6127–34.

Davidson-Pilon, Cameron. 2021. *CamDavidsonPilon/Lifelines*.

Devriendt, Floris, Jeroen Berrevoets, and Wouter Verbeke. 2019. “Why You Should Stop Predicting Customer Churn and Start Using Uplift Models.” *Information Sciences*.

Edward, Manoj and Sunil Sahadev. 2011. “Role of Switching Costs in the Service Quality, Perceived Value, Customer Satisfaction and Customer Retention Linkage.” *Asia Pacific Journal of Marketing and Logistics* 23(3):327–45.

Ehrlinger, John. 2016. “ggRandomForests: Exploring Random Forest Survival.” *arXiv:1612.08974 [Stat]*.

Fornell, Claes and Birger Wernerfelt. 1987. “Defensive Marketing Strategy by Customer Complaint Management: A Theoretical Analysis.” *Journal of Marketing Research* 24(4):337–46.

Fotso, Stephane and others. 2019. *PySurvival: Open Source Package for Survival Analysis Modeling*.

García, David L., Angela Nebot, and Alfredo Vellido. 2017. "Intelligent Data Analysis Approaches to Churn as a Business Problem: A Survey." *Knowledge and Information Systems* 51(3):719–74.

Gök, Mehmet, Tansel Özyer, and Jamal Jida. 2015. "A Case Study for the Churn Prediction in Turksat Internet Service Subscription." Pp. 1220–24 in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*. Paris, France: ACM Press.

Gupta, Sunil, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. 2006. "Modeling Customer Lifetime Value." *Journal of Service Research* 9(2):139–55.

Hung, Shin-Yuan, David C. Yen, and Hsiu-Yu Wang. 2006. "Applying Data Mining to Telecom Churn Management." *Expert Systems with Applications* 31(3):515–24.

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2(3):841–60.

Jafari-Marandi, Ruholla, Joshua Denton, Adnan Idris, Brian K. Smith, and Abbas Keramati. 2020. "Optimum Profit-Driven Churn Decision Making: Innovative Artificial Neural Networks in Telecom Industry." *Neural Computing and Applications* 32(18):14929–62.

Mellens, Martin, Marnik Dekimpe, and JBEM Steenkamp. 1996. "A Review of Brand-Loyalty Measures in Marketing." *Review of Business and Economic Literature* 41(4):507–33.

Perianez, Africa, Alain Saas, Anna Guitart, and Colin Magne. 2016. "Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles." Pp. 564–73 in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Montreal, QC, Canada: IEEE.

Reichheld, Frederick F. 1996. "Learning from Customer Defections." *Harvard Business Review* 74(2):56–67.

Risselada, Hans, Peter C. Verhoef, and Tammo H. A. Bijmolt. 2010. "Staying Power of Churn Prediction Models." *Journal of Interactive Marketing* 24(3):198–208.

Routh, Pallav, Arkajyoti Roy, and Jeff Meyer. 2020. "Estimating Customer Churn Under Competing Risks." *Journal of the Operational Research Society* 1–18.

Saunders, J. a. 1980. "Cluster Analysis for Market Segmentation." *European Journal of Marketing* 14(7):422–35.

Schober, Patrick and Thomas R. Vetter. 2018. "Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare." *Anesthesia and Analgesia* 127(3):792–98.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6(2):461–64.

Scrucca, Luca, Michael Fop, T. ,Brendan Murphy, and Adrian,E. Raftery. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R*

Journal 8(1):289.

Singer, Judith D. and John B. Willett. 1993. "It's about Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events." *Journal of Educational Statistics* 18(2):155–95.

Stensrud, Mats J. and Miguel A. Hernán. 2020. "Why Test for Proportional Hazards?" *JAMA* 323(14):1401–2.

Vijaya, J. and E. Sivasankar. 2019. "An Efficient System for Customer Churn Prediction Through Particle Swarm Optimization Based Feature Selection Model with Simulated Annealing." *Cluster Computing* 22(S5):10757–68.

Wang, Ping, Yan Li, and Chandan K. Reddy. 2017. "Machine Learning for Survival Analysis: A Survey." *arXiv:1708.04649 [Cs, Stat]*.

Appendix: Chunk options

7.1 Software versioning R

7.1.1 R

```r
cat(paste("#", capture.output(sessionInfo()), "\n", collapse  = ""))
```

```
## # R version 4.1.3 (2022-03-10)
## # Platform: x86_64-w64-mingw32/x64 (64-bit)
## # Running under: Windows 10 x64 (build 22622)
## #
## # Matrix products: default
## #
## # locale:
## # [1] LC_COLLATE=Portuguese_Portugal.1252  LC_CTYPE=Portuguese_Portugal.1252
## # [3] LC_MONETARY=Portuguese_Portugal.1252 LC_NUMERIC=C
## # [5] LC_TIME=Portuguese_Portugal.1252
## #
## # attached base packages:
## # [1] stats     graphics  grDevices utils     datasets  methods   base
## #
## # other attached packages:
## #  [1] mclust_5.4.10    labelled_2.9.1   kableExtra_1.3.4 gtsummary_1.6.0
## #  [5] visdat_0.5.3     readxl_1.4.0     stargazer_5.2.3  reticulate_1.25
## #  [9] ggplot2_3.3.6    dlookr_0.5.6     dplyr_1.0.9
## #
## # loaded via a namespace (and not attached):
## #  [1] reactable_0.2.3     webshot_0.5.3       httr_1.4.3
## #  [4] tools_4.1.3         utf8_1.2.2          R6_2.5.1
## #  [7] rpart_4.1.16        colorspace_2.0-3    withr_2.5.0
## # [10] tidyselect_1.1.2    gridExtra_2.3       curl_4.3.2
## # [13] compiler_4.1.3      extrafontdb_1.0     cli_3.3.0
## # [16] rvest_1.0.2         gt_0.6.0            xml2_1.3.3
## # [19] labeling_0.4.2      bookdown_0.27       scales_1.2.0
## # [22] mvtnorm_1.1-3       rappdirs_0.3.3      systemfonts_1.0.4
## # [25] stringr_1.4.0       digest_0.6.29       rmarkdown_2.14
## # [28] svglite_2.1.0       pkgconfig_2.0.3     htmltools_0.5.2
## # [31] showtext_0.9-5      extrafont_0.18      fastmap_1.1.0
## # [34] highr_0.9           htmlwidgets_1.5.4   rlang_1.0.2
## # [37] rstudioapi_0.13     sysfonts_0.8.8      shiny_1.7.1
## # [40] generics_0.1.2      farver_2.1.0        jsonlite_1.8.0
## # [43] magrittr_2.0.3      Formula_1.2-4       Matrix_1.4-1
## # [46] Rcpp_1.0.8.3        munsell_0.5.0       fansi_1.0.3
## # [49] gdtools_0.2.4       partykit_1.2-15     lifecycle_1.0.1
## # [52] stringi_1.7.6       yaml_2.3.5          inum_1.0-4
## # [55] grid_4.1.3          hrbrthemes_0.8.0    promises_1.2.0.1
## # [58] forcats_0.5.1       crayon_1.5.1        lattice_0.20-45
## # [61] haven_2.5.0         splines_4.1.3       hms_1.1.1
## # [64] knitr_1.39          pillar_1.7.0        glue_1.6.2
## # [67] evaluate_0.15       pagedown_0.18       broom.helpers_1.7.0
## # [70] vctrs_0.4.1         png_0.1-7           httpuv_1.6.5
## # [73] Rttf2pt1_1.3.10     cellranger_1.1.0    gtable_0.3.0
## # [76] purrr_0.3.4         tidyr_1.2.0         xfun_0.31
## # [79] mime_0.12           libcoin_1.0-9       xtable_1.8-4
## # [82] later_1.3.0         survival_3.3-1      viridisLite_0.4.0
## # [85] tibble_3.1.7        showtextdb_3.0      ellipsis_0.3.2
```

```
# or use message() instead of cat()
```