Customer dropout membership*

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

Extremadura University

22 September, 2021

Abstract

Abstract of the article. Here we can place more info.

Introduction

Research idea:

Context: An organization membership located in Portugal. The organization offers an annual membership for the members, the service subscription has several payment options:

- Men with a annual fee of 10€
- Women annual fee of 6€
- Correspondent fee 6€
- Retired fee 5€

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

- Student fee 2.5€
- under-14 fee 1€

Churn dropout prediction is a problem being addressed supported in the idea that the customers database is the most valuable asset that the organizations possess (Athanassopoulos 2000), which requires determining customers that will attrite (Alboukaey, Joukhadar, and Ghneim 2020a). Dropout implies in contractual business that the customer needs to renew their contracts to continue its usage (Ascarza and Hardie 2013).

However, in contractual settings the customer dropout represents an explicit ending of a relationship more punitive 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)

```
If an organization can predict a possible dropout and
develop countermeasures to avoid desertion, they can avoid customer defections that
lead to a loss of money. Reichheld (1996) evidenced that reducing dropout rates by 5%
(e.g., from 15% to 10% per year) could represent an increase in prots up to double, as
acquiring new customers costs 5 to 6 times more than retaining existing ones (Bhattachar
1998). Existing organizations are addressing this problem by shifting their target from
capturing new customers to preserving existing ones (García et al., 2017),
as investments in retention strategies have higher returns than acquisitions (Coussement
& Van den Poel, 2009). The importance of customer retention to maintain organizational p
et al., 2019) leads to the problem of how to quantify the nancial impact of customer ret
under the assumption that the organization goal should be related to the increase the li
customer to increase their prots. The customer lifetime value (CLV) allows us to measure
```

1. Address the global problem of customer dropout
2. The identification of approaches to predict dropout requires more than only, address the prediction accuracy such as … place existing studies addressing this…
3. A lot of effort has been placed testing the accuracy of existing algorithms, in this study we try to fill this gap and explore also a balancing between the model interpretability, accuracy, and the investment required.
4. Dropout prediction problem related to the timings

The approaches normally employing 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, Joukhadar, and Ghneim 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.

In this study, we adopt random survival forests which have never been used in understanding factors affecting membership in a sport club using existing data in a Sport Club. The analysis is based on the use of random survival forests in the presence of covariates that do not necessarily satisfy the PH assumption. Additionly we also propose a new approach combining clusters with survival analysis.

??? Add interpretability layer

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.

Methodology

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 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).

Dataset

Table @ref(tab:summarytable) shows data's summary statistics. The average age is 27.3 ± 20.1, the members have an attendance of 27 ± 45.8 with a membership of 11 ± 10.9 years.

Figure @ref(fig:membershipyear) shows the distribution of the dropout considering the number of years of membership.

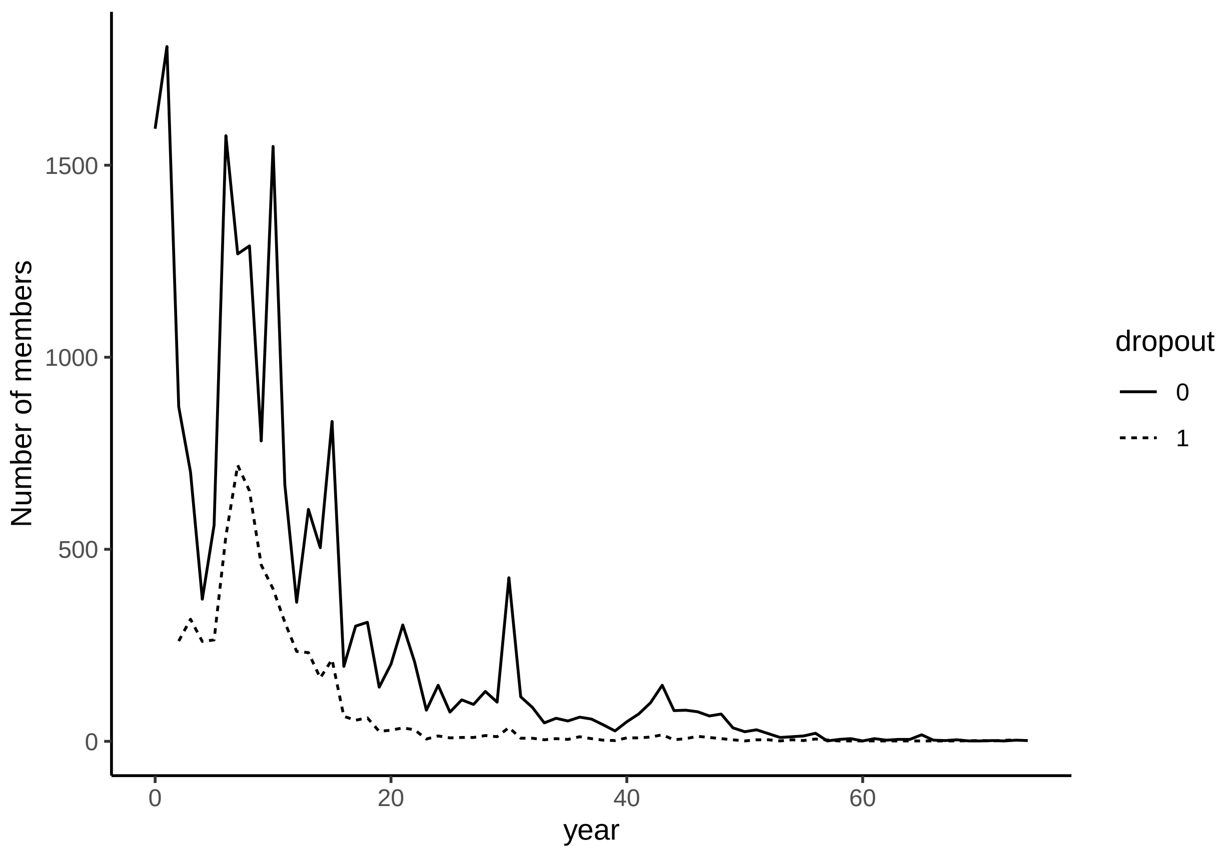


Figure 1: Number of members by year

Table 1: Summary statistics of features used

Characteristic	N = 25,316
Age in years, Mean (SD)	27 (20)
Male or female, %	
F	32%
M	68%
Single, married and other., %	
casado	20%
nao definido	30%
outro	2.0%
solteiro	48%
monthly_fee, %	
0	<0.1%
1	32%
2.5	28%
5	3.4%
6	12%
10	24%
total_amount, Mean (SD)	316 (494)
total_matches, Mean (SD)	27 (46)
season_matches, Mean (SD)	2.2 (4.1)
months_since_last_payment, Mean (SD)	19 (32)
dropout, %	22%
years_membership, Mean (SD)	11 (11)
stadium_access, %	40%
quart_stadium_entries, %	
1 a 21	10%
21 a 56	9.8%
56 a 105	10.0%
ate 1	60%
mais 105	10.0%
inscription_month, Mean (SD)	6.9 (3.4)

Model construction

Address the model construction... the categorical variables *sex*, *marital_status* and *quart_stadium_entries* where converted to dummy variables.

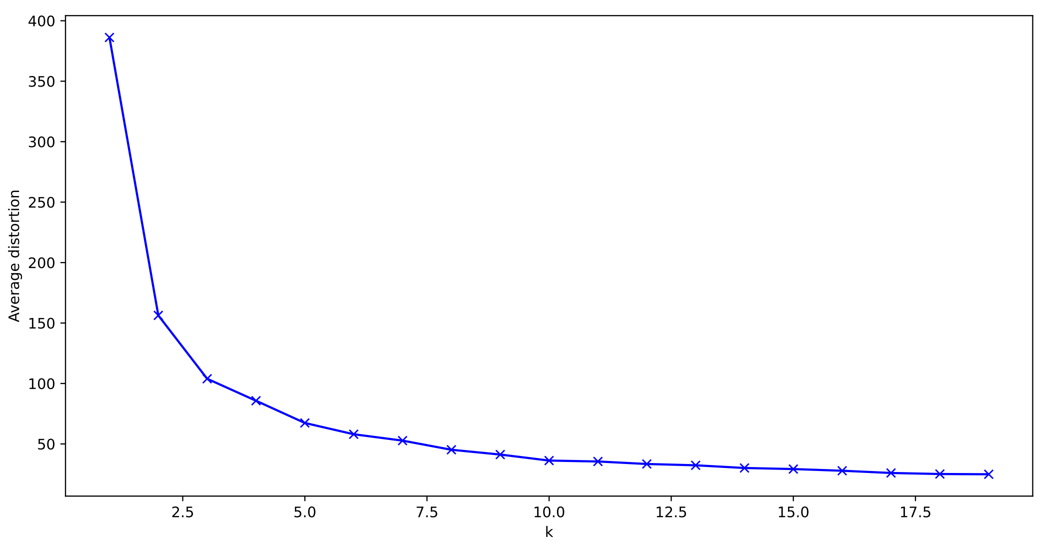
The random survival forest was developed using the package PySurvival (Fotso and others 2019). The most relevant variables predicting the dropout are analysed using the log-rank test. The metric variables are transformed to categorical using the quartiles to provide a statistical comparison of groups. The survival analysis was conducted using the package Lifelines (Davidson-Pilon 2021).

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. It is built upon the most commonly used machine learning packages such NumPy, SciPy and PyTorch. PySurvival is compatible with Python 2.7-3.7

Survival trees based model

In this model... The survival trees based model uses pysurvival random forest

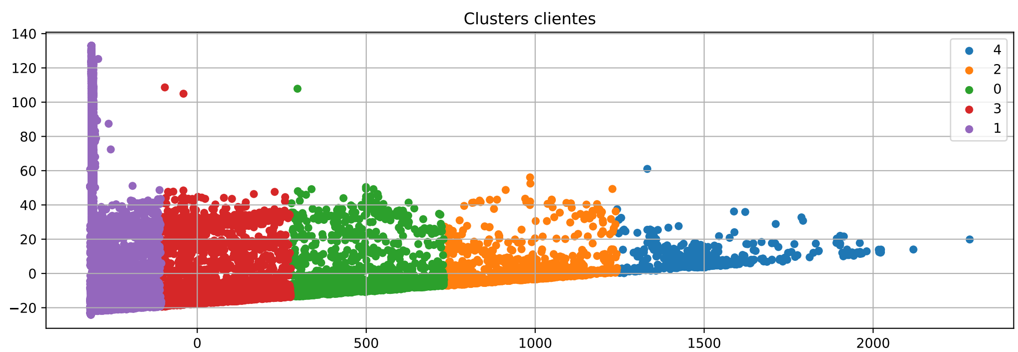
Removed the variables with greater correlations *total_matches* and *quart_stadium_entries*




```
##         age  ...  quart_stadium_entries_mais 105
## 0      83.0  ...                               0
## 1      88.0  ...                               0
## 2      73.0  ...                               0
## 3      97.0  ...                               0
```

```
## 4       97.0  ...                              0
## ...      ...  ...                            ...
## 25311     7.0  ...                              0
## 25312     8.0  ...                              0
## 25313     2.0  ...                              0
## 25314    14.0  ...                              0
## 25315    28.0  ...                              0
##
## [25316 rows x 14 columns]
```

```
## RandomSurvivalForestModel
```



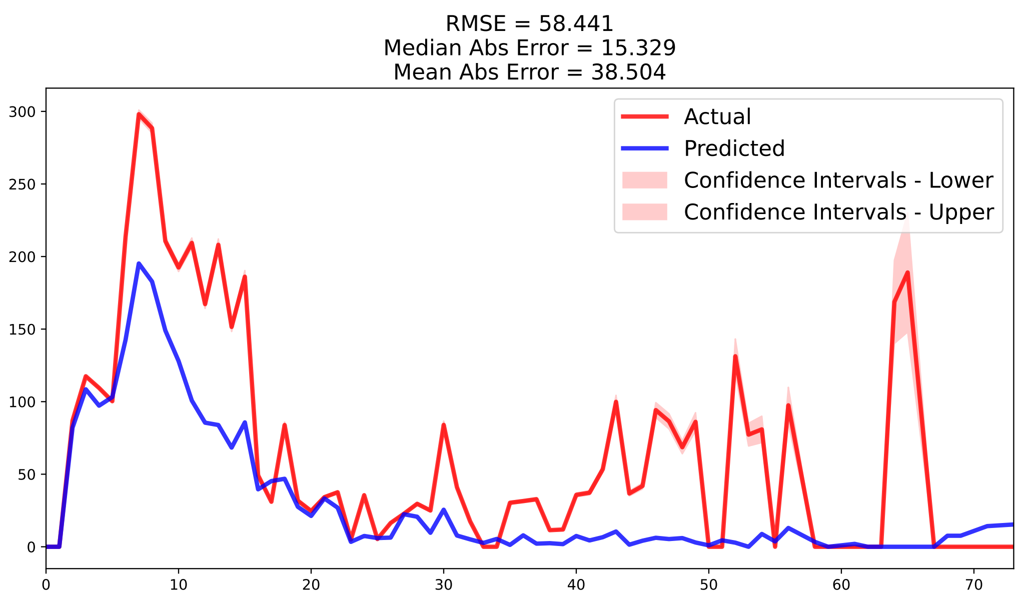


Table @ref(tab:summarytable2) shows variables importance.

Model building The model was built with with 70% of the data for training and 30% for testing. The survival model parameters where:

Table 2: Summary statistics of features used

feature	importance	pct_importance
dropout	17.1709170	0.2937936
months_since_last_payment	13.7848413	0.2358580
total_amount	6.7674785	0.1157912
season_matches	5.0909413	0.0871058
inscription_month	4.1582329	0.0711472
stadium_access	2.3607581	0.0403925
marital_status_solteiro	1.8905029	0.0323464
quart_stadium_entries_21 a 56	1.8425100	0.0315253
quart_stadium_entries_mais 105	1.6170003	0.0276668
sex_M	1.5765357	0.0269745
marital_status_outro	0.8011876	0.0137083
monthly_fee	0.7854033	0.0134382
age	0.5754982	0.0098467
marital_status_nao definido	0.0237108	0.0004057
quart_stadium_entries_56 a 105	-0.0060609	0.0000000
years_membership	-19.4149405	0.0000000

The model accuracy is very high in the first years. The prediction is very similar to the actual value. The absolute error mean of 39 customers.

Survival trees based model with clusters

Here we are will create clusters and developed the optimization within each cluster…

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).

In multivariate models are available the following approaches:

- “EII”spherical, equal volume
- “EEE”ellipsoidal, equal volume, shape, and orientation
- “VII”spherical, unequal volume
- “VVV”ellipsoidal, varying volume, shape, and orientation, which is used as default for initialization of EM algorithm

- VVI": diagonal, varying volume and shape t

Estou a ter problemas com o cálculo dos clusters com o BIC… estava a aqui a confirmar os clusters, talvez seja melhor reduzir as variáveis… testar a abordagem aos clusters no artigo dos vinhos…

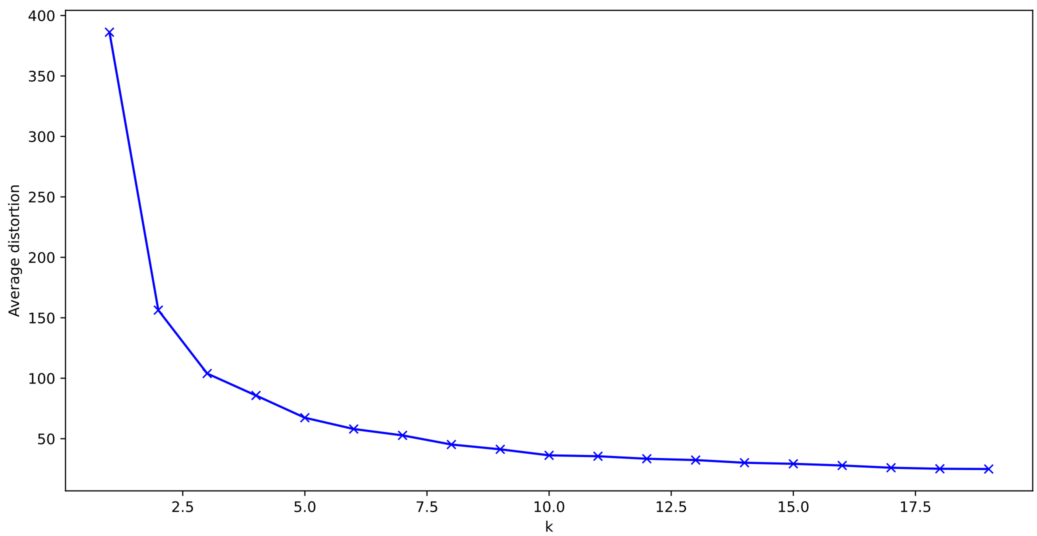
```
library(NbClust)
nb <- NbClust(y, diss=NULL, distance = "euclidean",
              min.nc=2, max.nc=5, method = "kmeans",
              index = "all", alphaBeale = 0.1)
hist(nb$Best.nc[1,], breaks = max(na.omit(nb$Best.nc[1,])))
```

```
## KMeans(n_clusters=1)
## KMeans(n_clusters=2)
## KMeans(n_clusters=3)
## KMeans(n_clusters=4)
## KMeans(n_clusters=5)
## KMeans(n_clusters=6)
## KMeans(n_clusters=7)
## KMeans()
## KMeans(n_clusters=9)
## KMeans(n_clusters=10)
## KMeans(n_clusters=11)
## KMeans(n_clusters=12)
## KMeans(n_clusters=13)
## KMeans(n_clusters=14)
## KMeans(n_clusters=15)
## KMeans(n_clusters=16)
## KMeans(n_clusters=17)
## KMeans(n_clusters=18)
## KMeans(n_clusters=19)
```

```
## [<matplotlib.lines.Line2D object at 0x7f5fd036c410>]
```

```
## Text(0.5, 0, 'k')
```

```
## Text(0, 0.5, 'Average distortion')
```



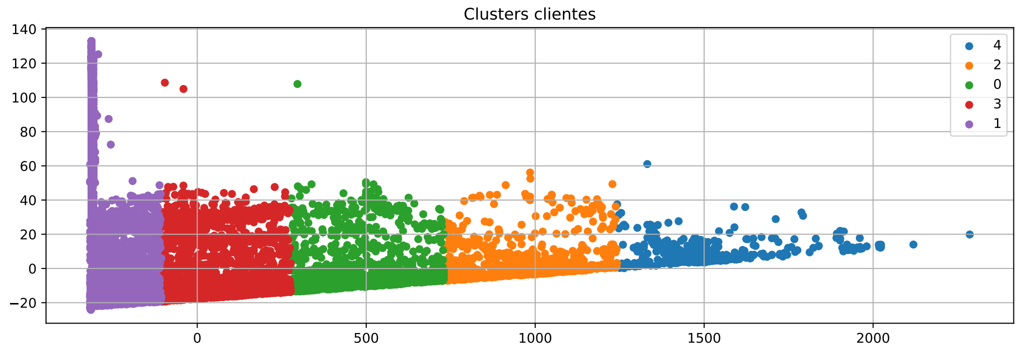
We are going to consider five clusters

```
## KMeans(n_clusters=5)
```

```
## 1    17070
## 3     2817
## 0     2419
## 2     2080
## 4      930
## Name: cluster, dtype: int64
```

```
## <matplotlib.collections.PathCollection object at 0x7f5fd0357e10>
## <matplotlib.collections.PathCollection object at 0x7f5fd0357d50>
## <matplotlib.collections.PathCollection object at 0x7f5fd036cb50>
## <matplotlib.collections.PathCollection object at 0x7f5fd02e7510>
## <matplotlib.collections.PathCollection object at 0x7f5fd02e79d0>
```

```
## <matplotlib.legend.Legend object at 0x7f5fd03a08d0>
```



TODO: Estava aqui… fit the model for the five clusters… compare the performance with the model without clusters… corrigir acima para selecionar só as features.. Explorart-SNE for better visualization…

Open questions - to remove

- RQ1: What is the current state of the research being developed?
- RQ2: What algorithms have been used to predict dropout?
- RQ3: What are the features used to predict dropout?
- RQ4: When does dropout occur?
- RQ5: How is the accuracy of the machine learning algorithms in predicting dropout measured?

From RQ1, it was possible to identify some business areas that are under-researched, such as the energy sector, education, logistics and hospitality. Compared to other business areas such telecom or the financial sector, research on the energy sector or water supply is lacking, considering the contractual settings that are assumed to provide such types of services. Considering the business model of many software companies as software as a services (SaS), the number of research works is also surprisingly low.

RQ2 also provided an overall perspective related to the algorithms being used to predict customer dropout. The first could be the importance and wider adoption of decision trees and random forests (Antipov and Pokryshevskaya 2010; Benoit and Van den Poel 2012; Burez and Van den Poel 2007), and logistic regression (Coussement, Benoit, and Van den Poel 2010), which could be due to its higher interpretability and flexibility (Keramati et al. 2014). Interpretability is an important aspect for the marketing department in the extraction of valuable information from the model to develop effective retention strategies (Verbeke et al. 2012). The problem arises in the balancing between interpretability and the higher performance of the algorithms inspired by nature (such as neural networks). From a business perspective, dropout prediction should also be considered as a business objective, which requires more than predicting if the customer will churn or not (Devriendt, Berrevoets, and Verbeke 2019), where higher interpretability provides better support in the development of retention strategies. The developed SLR also raises the possibility of integrating different algorithms using ensemble methods or integrating several models using a hybrid approach. None of the studies integrated the survival approach to predict customer dropout, for example, using a hybrid approach.

It is considered positive if actions are developed to retain customers, but the problems should also be considered, such as the following: (1) customers who have greater risk of dropout should be targeted to provide a base for a better ROI in the retention strategies (Coussement and Van den Poel 2008; Xie et al. 2009) and (2) the retention strategies should be developed focused on customers with higher satisfaction, or its inclusion could be a reminder of the contractual agreement nearing an end and could lead to churn (Devriendt, Berrevoets, and Verbeke 2019).

From RQ3, several types of features being used were able to be identified, such as demographic, behavioral, and economic indicators, pictorial data, network relationships or high cardinality features. The problem that arises is that some studies used data and features that were not described, and this creates a major issue, How can reproducibility be developed in a study without the availability of the data or the identification of the features used? Considering that science is driven by data, with the development of new technologies, the increasing complexity of research and the amount of data collected, the challenge is to ensure that research is available to all (Hanson, Sugden, and Alberts 2011); this requires both availability of the data and the algorithms so that they can be explored by other researchers. The features are selected mainly to verify the performance of the models, and are essential to performance prediction, accuracy, and the steps for processing the data, which are fundamental to improve the model accuracy (Azeem, Usman, and Fong 2017).

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, Joukhadar, and Ghneim 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.

According to each business model, the timeframe could be addressed considering the survival probability according to the customer relationship age, and dropout predictions could be developed according to these survival probabilities, as suggested by Esteves and Mendes-Moreira (2016), to investigate which data timeframe produces the best result and how the efficiency of the models is influenced by this timeframe. Exploring the duration of the relation and the understanding of the features that increase or decrease that duration seems to be an important approach that could complement the existing approaches to predicting dropout.

From RQ5, the literature analysis showed that different types of questions arise. Which are the best approaches to develop the analysis of the performance in predicting dropout? Several metrics are customer dropout is to improve the performance of organizations in retaining customers, which is a management problem in which data mining is adopted (Verbeke et al. 2012). The goals of the model should be formulated considering the context of the problem that is being addressed; in marketing retention strategies, the up-lift supports the development of proactive actions to minimize the investment in retention strategies (Coussement and Van den Poel 2008). Some assumptions that underlie the adoption of uplift metrics consider that customers with a higher risk of churning could not be the best targets, as suggested by Ascarza (2018). Other researchers addressed the problem using the top-decile lift to develop more proactive actions to retain the customers at risk of churning [Coussement and Van den Poel (2008);xie_customer_2009]. This approach considers the 10% of customers with more risk, and investments in retention strategies should be developed that distinguish the churners susceptible to marketing actions from those who will leave anyway (Coussement, Lessmann, and Verstraeten 2017). Although uplift models seem to be good strategies, they should also used, such as AUC, sensitivity, specificity, recall, precision, and F-score. However, the goal ofconsider factors other than risk and customer satisfaction, as not taking this into consideration could be counterproductive and the model should be removed from the retention strategy.

The true business objective is to reduce customer churn. Customers who are about to churn but cannot be retained should be excluded from the campaign, as targeting them will be a waste of scarce resources (Devriendt, Berrevoets, and Verbeke 2019). Using these models seem to be a good strategy, as they can outperform predictive models that consider only accuracy from a profitability busshould be considered that customers with a higher risk of churning may not be the best targets to develop retention strategies. Those perspectives entail the dropout.

that a business context, or the clarification of a business objective underlying the prediction of customer dropout, should be developed, to clarify which objectives should be achieved before employing the profitability of reducing g machine learning algorithms. Surprisingly, the analyzed studies did not address the customer lifetime value as an objective to optimize considerininess perspective.

Aspects to consider

- Interpretability from RQ2
- The business objective is to increase the number of members and organization profits
- piping several algorithms to improve accuracy. Aka hybrid approach
- Alboukaey, Joukhadar, and Ghneim (2020b) proposes …
- grep the articles addressing hybrid: pdfgrep -ri "hybrid.{1,10} approach"

Results

In this section, we present our experiments to validate the proposed models, comparing against other approaches. The models where optimized using the hyper-parameters Grid Search technique. The explored hyper-parameters and the best values of these parameters for every model are listed in (ref:table1).

(ref:table1) Table 2

Model name	Explored parameters values	Best parameters
Survival trees	pysurvival random forest	a
Survival trees with clusters	pysurvival random forest with clusters	a
Scikit survival trees	scikit survival	a
Scikit survival with clusters	scikit with clusters	a
Scikit survival gradient boost	scikit survival gradient boost	a
Scikit survival gradient boost with clusters	scikit with clusters	a

Table 3: Hyper-parameters best values

Conclusion

Article Ascarza

- Retention Futility: Targeting High-Risk Customers Might be Ineffective (Ascarza 2018)

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. Journal of Marketing Research, 55(1), 80-98. sim. https://doi.org/10.1509/jmr.16.0163

Example of Developed actions:

```
Each month, the company identified the customers who were up for renewal and
split them (randomly and evenly) between a treatment group that received a
"thank you" gift with the letter and a control group that received only the
renewal latter.
```

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.

Antipov, Evgeny and Elena Pokryshevskaya. 2010. "Applying CHAID for Logistic Regression Diagnostics and Classification Accuracy Improvement." *Journal of Targeting, Measurement and Analysis for Marketing* 18(2):109–17.

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.

Azeem, Muhammad, Muhammad Usman, and A. C. M. Fong. 2017. "A Churn Prediction Model for Prepaid Customers in Telecom Using Fuzzy Classifiers." *Telecommunication Systems* 66(4):603–14.

Benoit, Dries F. and Dirk Van den Poel. 2012. "Improving Customer Retention in Financial Services Using Kinship Network Information." *Expert Systems with Applications* 39(13):11435–42.

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.

Burez, Jonathan and Dirk Van den Poel. 2007. “CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services.” *Expert Systems with Applications* 32(2):277–88.

Coussement, Kristof, Dries F. Benoit, and Dirk Van den Poel. 2010. “Improved Marketing Decision Making in a Customer Churn Prediction Context Using Generalized Additive Models.” *Expert Systems with Applications* 37(3):2132–43.

Coussement, Kristof, Stefan Lessmann, and Geert Verstraeten. 2017. “A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry.” *Decision Support Systems* 95:27–36.

Coussement, Kristof and Dirk Van den Poel. 2008. “Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques.” *Expert Systems with Applications* 34(1):313–27.

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*.

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

Esteves, Georgina and Joao Mendes-Moreira. 2016. “Churn Perdiction in the Telecom Business.” Pp. 254–59 in *2016 Eleventh International Conference on Digital Information Management (ICDIM)*. Porto, Portugal: IEEE.

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.

Hanson, Brooks, Andrew Sugden, and Bruce Alberts. 2011. "Making Data Maximally Available." *Science (New York, N.Y.)* 331(6018):649.

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.

Keramati, A., R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi. 2014. "Improved Churn Prediction in Telecommunication Industry Using Data Mining Techniques." *Applied Soft Computing* 24:994–1012.

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.

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.

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.

Verbeke, Wouter, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. 2012. "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach." *European Journal of Operational Research* 218(1):211–29.

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]*.

Xie, Yaya, Xiu Li, E. W. T. Ngai, and Weiyun Ying. 2009. "Customer Churn Prediction Using Improved Balanced Random Forests." *Expert Systems with Applications* 36(3, Part 1):5445–49.

Appendix: Chunk options

Software versioning

R

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

```
## # R version 4.1.1 (2021-08-10)
## # Platform: x86_64-pc-linux-gnu (64-bit)
## # Running under: Ubuntu 20.04.3 LTS
## #
## # Matrix products: default
## # BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## # LAPACK: /home/sobreiro/miniconda3/envs/survival/lib/libmkl_intel_lp64.so
## #
## # locale:
## # [1] en_US.UTF8
## #
## # attached base packages:
## # [1] stats     graphics  grDevices utils     datasets  methods   base
## #
## # other attached packages:
## #  [1] labelled_2.8.0   kableExtra_1.3.4 gtsummary_1.4.2  visdat_0.5.3
## #  [5] readxl_1.3.1     stargazer_5.2.2  reticulate_1.20  ggplot2_3.3.5
## #  [9] dlookr_0.4.5     dplyr_1.0.7      rmarkdown_2.11   nvimcom_0.9-115
## #
## # loaded via a namespace (and not attached):
## #   [1] webshot_0.5.2       RColorBrewer_1.1-2  httr_1.4.2
## #   [4] tools_4.1.1         backports_1.2.1     utf8_1.2.2
## #   [7] R6_2.5.1            rpart_4.1-15        Hmisc_4.5-0
## #  [10] nortest_1.0-4       DBI_1.1.1           colorspace_2.0-2
## #  [13] nnet_7.3-16         withr_2.4.2         tidyselect_1.1.1
## #  [16] gridExtra_2.3       curl_4.3.2          compiler_4.1.1
```

```
## #  [19] extrafontdb_1.0     cli_3.0.1           rvest_1.0.1
## #  [22] gt_0.3.0            htmlTable_2.2.1     xml2_1.3.2
## #  [25] sandwich_3.0-1      labeling_0.4.2      scales_1.1.1
## #  [28] checkmate_2.0.0     mvtnorm_1.1-2       proxy_0.4-26
## #  [31] RcmdrMisc_2.7-1     rappdirs_0.3.3      systemfonts_1.0.2
## #  [34] stringr_1.4.0       digest_0.6.27       foreign_0.8-81
## #  [37] svglite_2.0.0       rio_0.5.27          base64enc_0.1-3
## #  [40] jpeg_0.1-8.1        pkgconfig_2.0.3     htmltools_0.5.2
## #  [43] extrafont_0.17      highr_0.9           fastmap_1.1.0
## #  [46] htmlwidgets_1.5.3   rlang_0.4.11        rstudioapi_0.13
## #  [49] prettydoc_0.4.1     farver_2.1.0        generics_0.1.0
## #  [52] jsonlite_1.7.2      zoo_1.8-9           zip_2.2.0
## #  [55] car_3.0-11          magrittr_2.0.1      Formula_1.2-4
## #  [58] Matrix_1.3-4        Rcpp_1.0.7          munsell_0.5.0
## #  [61] fansi_0.5.0         abind_1.4-5         gdtools_0.2.3
## #  [64] partykit_1.2-13     lifecycle_1.0.0     stringi_1.7.4
## #  [67] yaml_2.2.1          inum_1.0-4          carData_3.0-4
## #  [70] MASS_7.3-54         grid_4.1.1          hrbrthemes_0.8.0
## #  [73] forcats_0.5.1       crayon_1.4.1        lattice_0.20-44
## #  [76] haven_2.4.3         splines_4.1.1       hms_1.1.0
## #  [79] knitr_1.33          pillar_1.6.2        glue_1.4.2
## #  [82] evaluate_0.14       latticeExtra_0.6-29 broom.helpers_1.3.0
## #  [85] data.table_1.14.0   png_0.1-7           vctrs_0.3.8
## #  [88] Rttf2pt1_1.3.8      cellranger_1.1.0    tidyr_1.1.3
## #  [91] gtable_0.3.0        purrr_0.3.4         assertthat_0.2.1
## #  [94] xfun_0.26           openxlsx_4.2.4      libcoin_1.0-8
## #  [97] e1071_1.7-7         class_7.3-19        survival_3.2-13
## # [100] viridisLite_0.4.0   tibble_3.1.4        cluster_2.1.2
## # [103] corrplot_0.90       ellipsis_0.3.2
```

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

Other used tools

- Visidata for quick exploratory. VisiData is a free, open-source tool that lets you quickly open, explore, summarize, and analyze datasets in your computer's terminal.