# Customer dropout membership

Pedro Sobreiro

27-06-2021

**Abstract**

Prediction of customer dropout with contractual settings

## Customer dropout membership :technologist: :moneybag: :chart_with_upwards_trend:

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€
- Student fee 2.5€
- under-14 fee 1€

## Methodology

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. 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. The random survival forest was developed using the package PySurvival (Fotso et al. 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).

## Packages installation

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

```
# create environment with python 3.7
conda create --name survival python=3.7
# activate environment
conda activate survival
```

```
# package essentials
conda install -c conda-forge jupyter
conda install -c conda-forge jupyterlab
conda install -c conda-forge xlrd
conda install -c conda-forge openpyxl
conda install -c conda-forge lifelines
# install PySurvival dependencies
conda install -c conda-forge numpy
conda install -c conda-forge scipy
conda install -c conda-forge scikit-learn
conda install -c conda-forge pytorch

# install c++ dependencies
sudo apt install gcc-8 g++-8
# edit .bashrc or .zshrc according the terminal used then source
# e.g. source ~/.zshrc
export CXX=/usr/bin/g++-8
export CC=/usr/bin/gcc-8
# install pysurvival after dependencies are resolved by conda
pip install pysurvival
```

```
@Misc{ pysurvival_cite,
  author = {Stephane Fotso and others},
  title = {{PySurvival}: Open source package for Survival Analysis modeling},
  year = {2019--},
  url = "https://www.pysurvival.io/"
}
```

# Running the model

```
from pysurvival.utils.display import correlation_matrix
correlation_matrix(df[features], figure_size=(10,10), text_fontsize=8)
```

### Removed the variables with greater correlations

```
to_remove = ['totalJogos', 'idaEstadio']
features = np.setdiff1d(features, to_remove).tolist()
```

### Model building

The model was built with with 60% of the data for training and 40% for testing. The survival model parameters where:

```
from pysurvival.models.survival_forest import RandomSurvivalForestModel
csf = RandomSurvivalForestModel(num_trees=200)
csf.fit(X_train, T_train, E_train, max_features='sqrt', max_depth=5, min_node_size=20)
```

The model accuracy is very high in the first years. The prediction is very similar to the actual value.

All the outputs are available here

# 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:

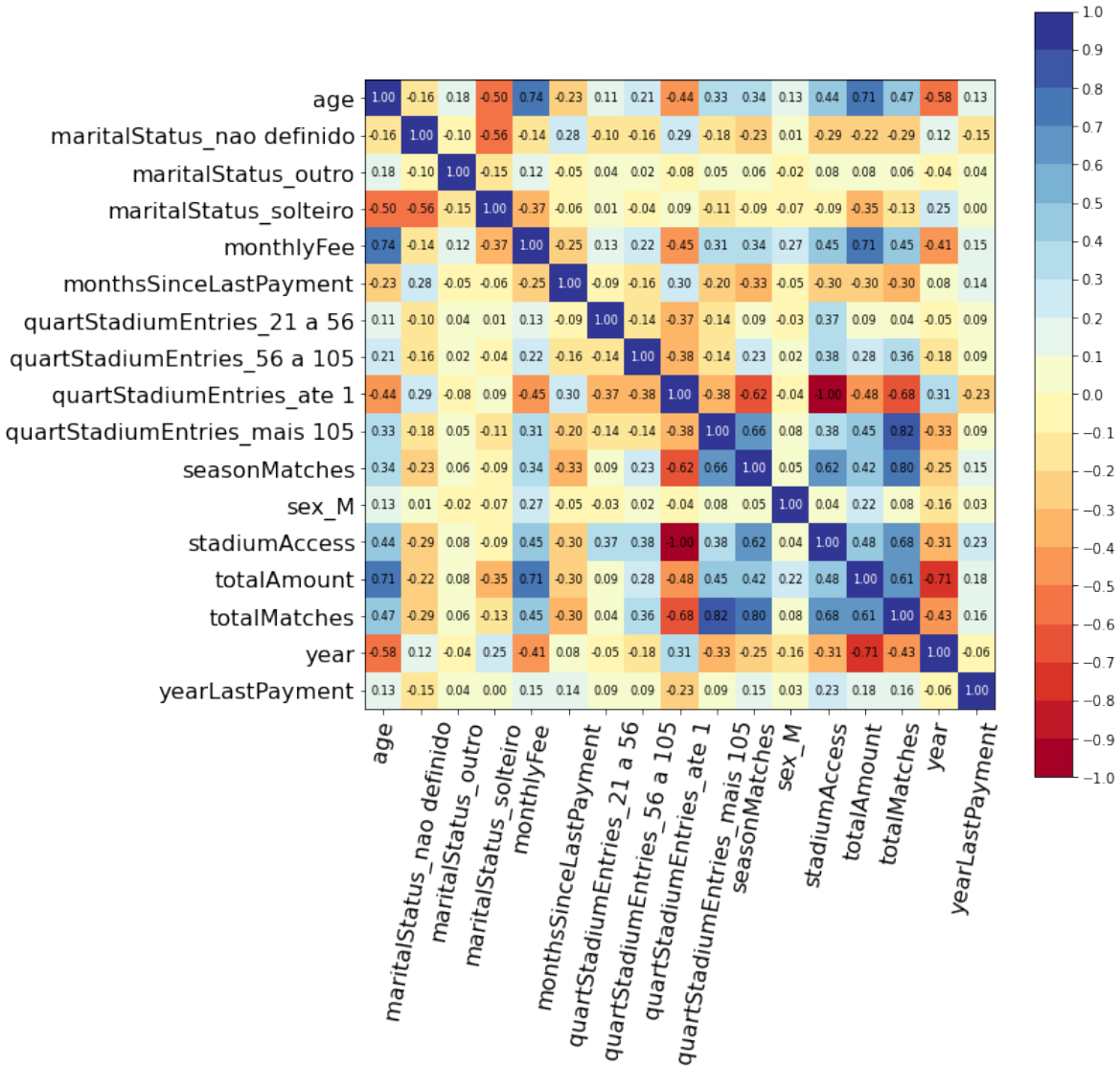


Figure 1: image

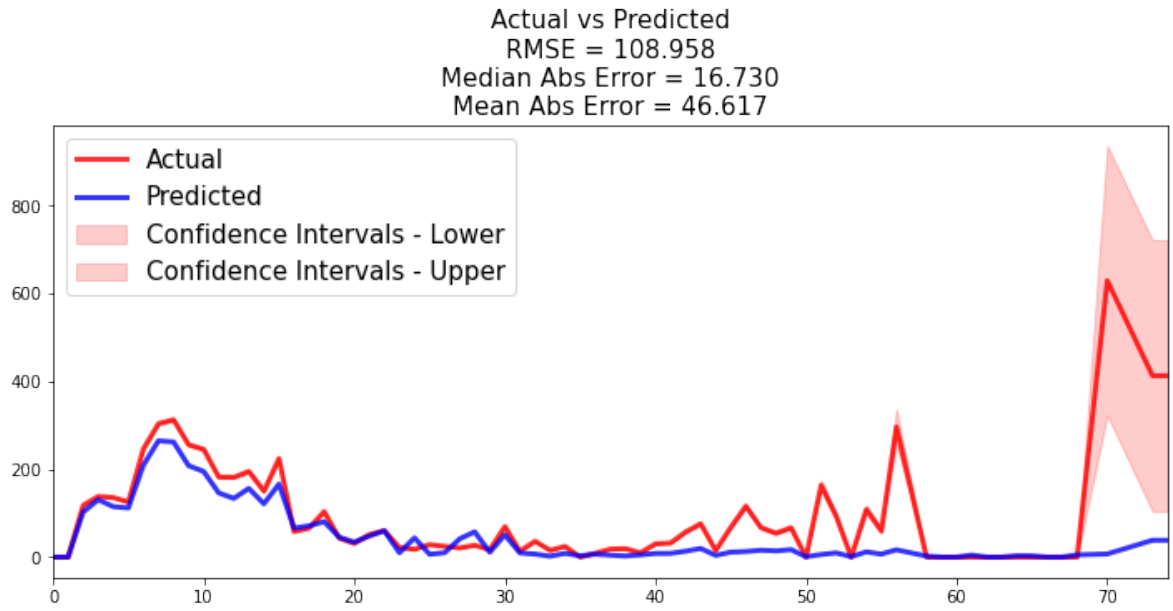


Figure 2: Prediction accuracy

```
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.
```

## Open questions

- 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; Jonathan 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 (Xie et al. 2009; Coussement and Van den Poel 2008) 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 2020b). 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 (Perianez et al. 2016; J. Burez and Vandenpoel 2008). 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 (2020a) proposes ...
- grep the articles addressing hybrid: pdfgrep -ri "hybrid.{1,10} approach"

## 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.

## Bibliography


Alboukaey, Nadia, Ammar Joukhadar, and Nada Ghneim. 2020a. "Dynamic Behavior Based Churn Prediction in Mobile Telecom." *Expert Systems with Applications* 162: 113779. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113779.

———. 2020b. "Dynamic Behavior Based Churn Prediction in Mobile Telecom." *Expert Systems with Applications* 162 (December): 113779. https://doi.org/10.1016/j.eswa.2020.113779.

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. https://doi.org/10.1057/jt.2010.3.

Ascarza, Eva. 2018. "Retention Futility: Targeting High-Risk Customers Might Be Ineffective." *Journal of Marketing Research* 55 (1): 80–98. https://doi.org/10.1509/jmr.16.0163.

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. https://doi.org/10.1007/s11235-017-0310-7.

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. https://doi.org/10.1016/j.eswa.2012.04.016.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

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. https://doi.org/10.1016/j.eswa.2005.11.037.

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. https://doi.org/10.1016/j.eswa.2007.07.036.

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. https://doi.org/10.1016/j.eswa.2009.07.029.

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 (March): 27–36. https://doi.org/10.1016/j.dss.2016.11.007.

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. https://doi.org/10.1016/j.eswa.2006.09.038.

Davidson-Pilon, Cameron. 2021. *CamDavidsonPilon/Lifelines*. https://github.com/CamDavidsonPilon/lifelines.

Devriendt, Floris, Jeroen Berrevoets, and Wouter Verbeke. 2019. "Why You Should Stop Predicting Customer Churn and Start Using Uplift Models." *Information Sciences*, December. https://doi.org/10.1016/j.ins.2019.12.075.

Ehrlinger, John. 2016. "ggRandomForests: Exploring Random Forest Survival." *arXiv:1612.08974 [Stat]*, December. http://arxiv.org/abs/1612.08974.

Esteves, Georgina, and Joao Mendes-Moreira. 2016. "Churn Perdiction in the Telecom Business." In *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, 254–59. Porto, Portugal: IEEE. https://doi.org/10.1109/ICDIM.2016.7829775.

Fotso, Stephane et al. 2019. *PySurvival: Open Source Package for Survival Analysis Modeling*. https://www.pysurvival.io/.

Hanson, Brooks, Andrew Sugden, and Bruce Alberts. 2011. "Making Data Maximally Available." *Science (New York, N.Y.)* 331 (6018): 649. https://doi.org/10.1126/science.1203354.

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 (November): 994–1012. https://doi.org/10.1016/j.asoc.2014.08.041.

Perianez, Africa, Alain Saas, Anna Guitart, and Colin Magne. 2016. "Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles." In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 564–73. Montreal, QC, Canada: IEEE. https://doi.org/10.1109/DSAA.2016.84.

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. https://doi.org/10.1016/j.ejor.2011.09.031.

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. https://doi.org/10.1016/j.eswa.2008.06.121.