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)
```

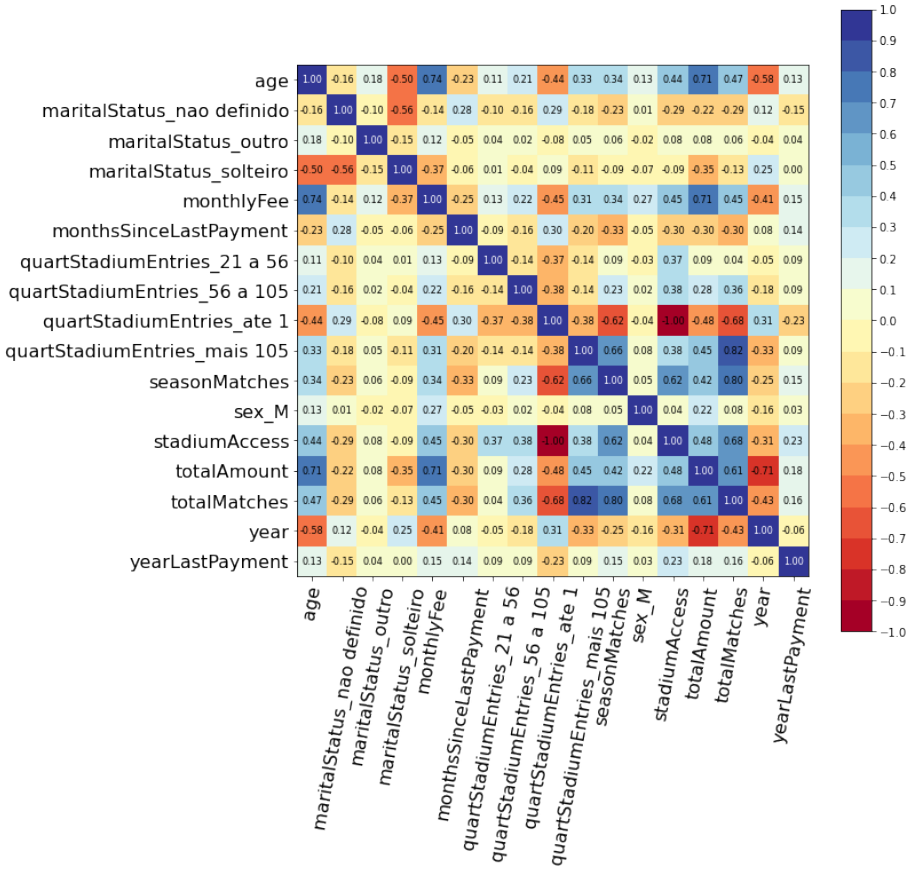

Figure 1: image

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.

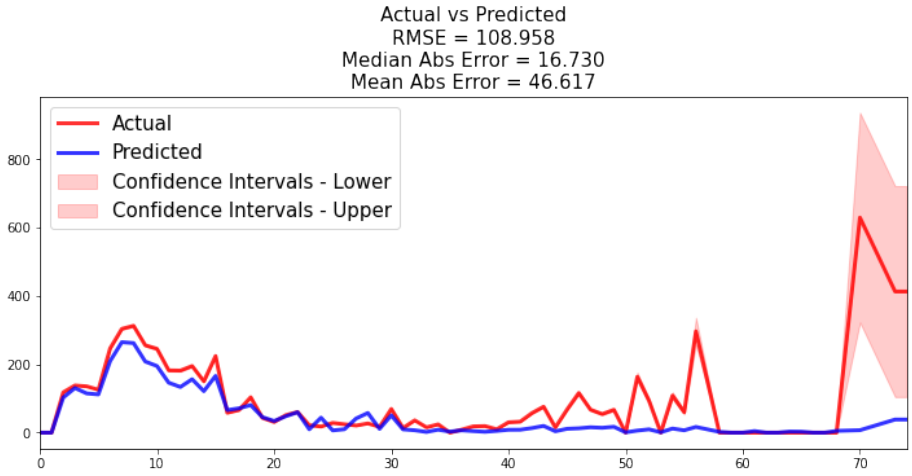

Figure 2: Prediction accuracy

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:

```
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.
```

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
-
```

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

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.

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

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

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

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