# SURVIVAL RATE ANALYSIS - DATA VALORIZATION FOR FINTECH

MSC. DATA SCIENCE AND ECONOMICS

**Gregorio Luigi Saporito**
Department of Economics, Management and Quantitative Methods
Department of Computer Science
University of Milan
Milan, Italy

**Tommaso Pessina**
Department of Economics, Management and Quantitative Methods
Department of Computer Science
University of Milan
Milan, Italy

October 26, 2020

**ABSTRACT**

The aim of this analysis is the one of determine if a person will conclude the subscription process to a virtual piggy bank application and/or where he is stuck and try to learn something from this kind of data, like the time to conclude the process, the percentage of people that will really conclude the process (with respect to those who are stuck in the process) and try to build a model for predict if a people will conclude or not the subscription (basing on his/her behavior). The data were pre-anonymized by the company that provided them.

***K*eywords** R · Statistics · Survival Analysis · Logistic Regression · More

## 1 Introduction

The original dataset provided by the company was organised as follow:

- ID: unique identifier for the event;
- COMPLETED_STEP: identify the completed step of the event in JSON format;
- DATE_EVENT: it contains the date and the time of the event;
- NETWORK_ID: represent the network by which the event occurred;
- USER_ID: unique identifier for the user;

Each step represent a different passage of the subscription process and they are:

- anagrafica-a, anagrfica-b, anagrafica-c, anagrfica-d: this are the step that ask to the client the personal data;
- codice fiscale: here the user should provide his/her fiscal code;
- antiriciclaggio: in this step is ask to the user if he/she is a politics or an "exposed" person and the source of his/her money;
- conclude: if a user arrives here he/she has concluded the subscription process;
- contract-activation, contract-subsritpion.

Starting from this dataset one can think about which kind of information it can deliver, but for see it we should run some operation over it. As matter of fact we have grouped users with their id in order to calculate all of this information. In more details, we create a new dataset which include:

- CLIENT_id: it is the unique identifier of the user;
- surv_time: we calculate the difference between the DATE_EVENT of the first step of a certain user and the DATE_EVENT of his last step in the process, and we see it as hour;
- status: is a flag that will assume value 1 if a user arrive through the step "conclude";
- extra_attempts: indicate if the user has done other attempt of subscription;
- main_extra_attempt: it will assume a value different from "none" if the previous item is different from zero and it represent the last step in which he/she is stuck;
- stuck: contain the step (of the subscription process) in which the user is stuck;
- mean_time: it is the mean time that the user spent in each of the steps that he/she has concluded.

All our analysis is based on this new dataset.

## 2 Survival analysis

survival

### 2.1 Kaplan-Meier estimates of the probability of survival over time

Kaplan-Meier

## 3 Predict the probability of conclude the subscription

logistic

## 4 Conclusion