SURVIVAL RATE ANALYSIS

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 27, 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

In this chapter we will discuss about the survival analysis with the aim to see how many people go all the way through the subscription process (along a time scale) or where they are stuck. Said that, we can also see if a person will conclude the subscription after a while of begin stuck.
First of all, we see that there are many outliers for survival time and few clients go through the system for a long time. After that, we can visually analyze in Figure 1 where the client are stuck in the subscription process.

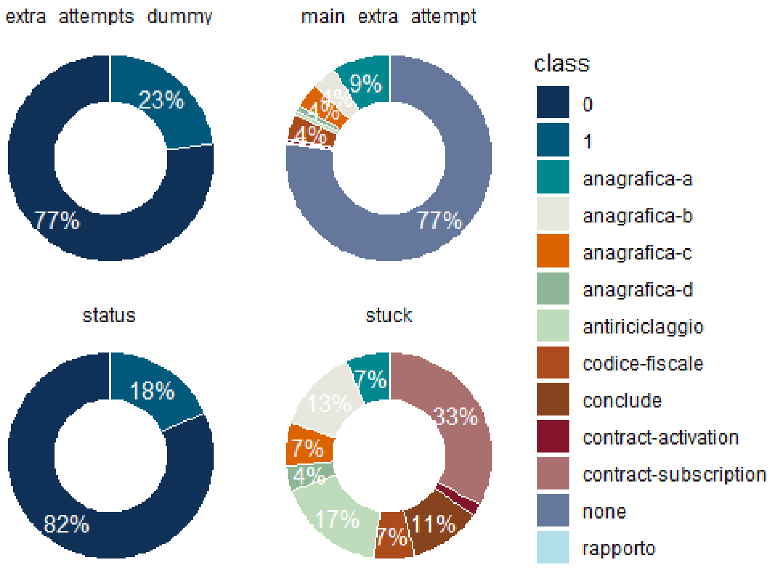

Figure 1: Donut graph of stuck users

Before discuss in detail our survival analysis, we shuold say few word about which kind of method we used. We used the "Kaplan–Meier estimator" that is a method used (especially in medical research) to estimate the probability that something will survive over time.
This is the basis of our analysis, that we will discuss in the next subchapter.

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

The first thing that we can say is that, as shown in Figure 2, as time passes the probability of concluding increases.

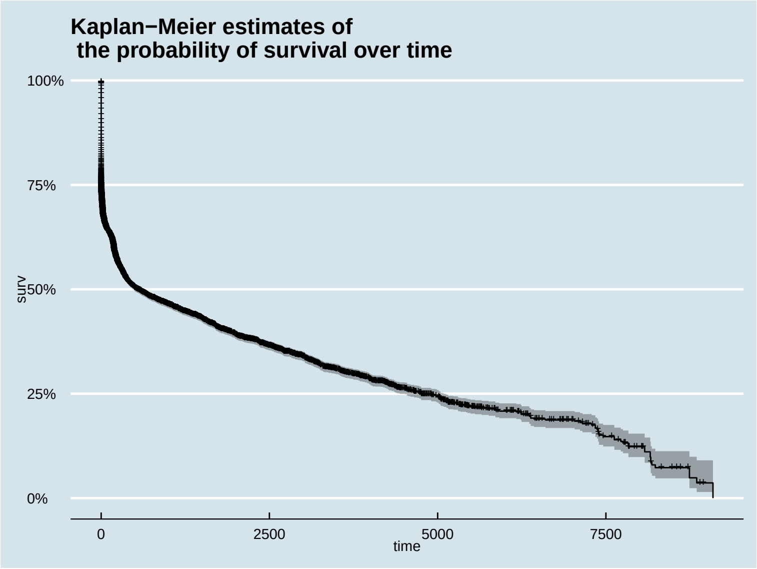

Figure 2: Probability of survival over time

Said that, we can now we look at the survival curve between who makes extra attempts and who does not. As shown in Figure 3, who makes extra attempts is more likely not to conclude and there are small overlaps between the curves.

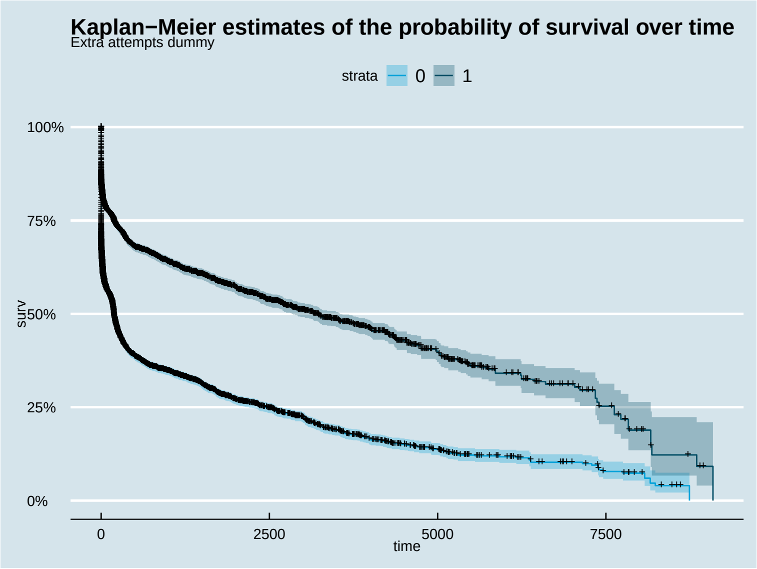

Figure 3: Probability of survival over time with extra attempt

Now, we can compare groups where the main stuck occurs and remove the ones with a too high confidence interval but there are still a lot of overlapps with confidence intervals though, as we can see in Figure 4.

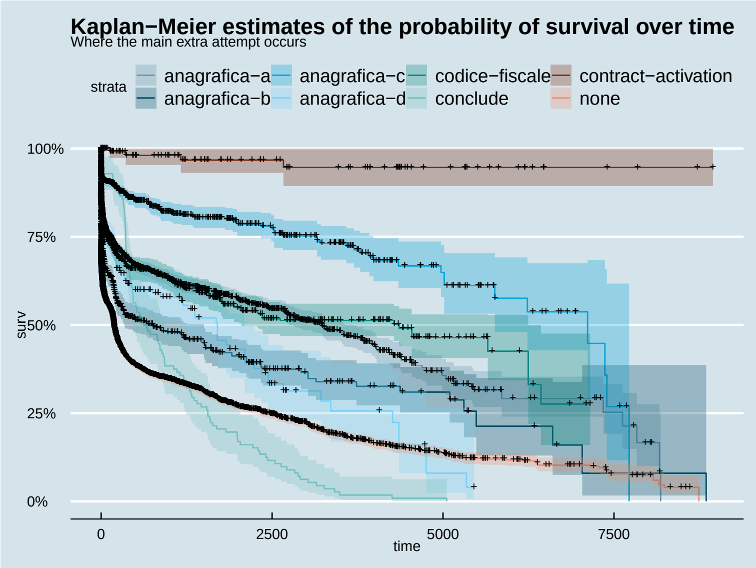

Figure 4: Total probability of survival over time with extra attempt

Finally, we look at where the client was last stuck in Figure 5.

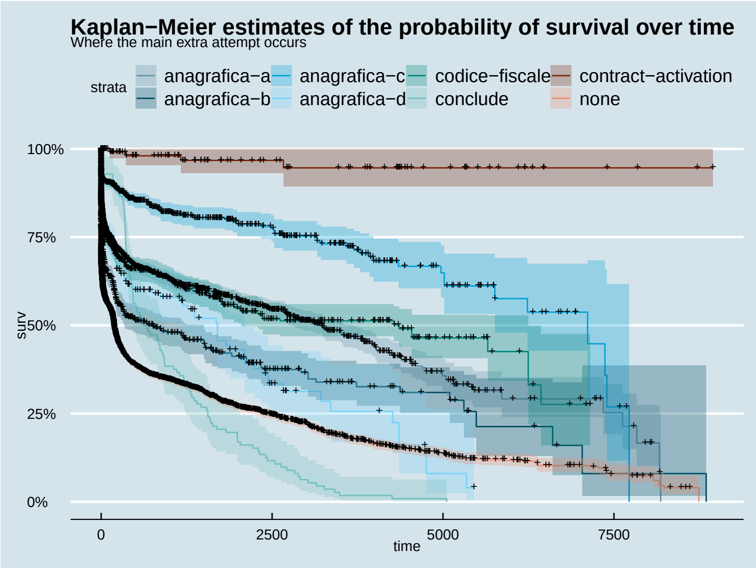

Figure 5: Probability of survival over time where the client is stuck

3 Predict the probability of conclude the subscription

Furthermore we can expand our analysis by creating a model for predicting the probability that a user will conclude the subscription process basing on his/her behavior.
Firstly we choose to use the Logistic Regression because is particular indicated in the categorical case, i.e. will or will not conclude the process. We start by dividing the dataset in train and test part with the rule of 7-3, moreover with divide this data by status (that can assume only value 0 or 1).
After having created this to sub-dataset (i.e. the train and the test set), we can define our model with "status" as response variable and "stuck", "surv_time", "extra_attempts" and "mean_time" as explanatory variable.
We can see below the summary of our model. As we can see, not all the variable are really explanatory for our estimate.
We can then compute the optimal score that minimizes the misclassification error for the above model, that is 0.95.

Listing 1: R output of the logist regression

```
Call:
glm(formula = status ~ stuck + surv_time + extra_attempts + mean_time,
    family = binomial(link = "logit"), data = trainingData)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-1.89525   -0.28149   -0.05391    0.00008    3.11563

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -5.144e+00  5.019e-01 -10.250  < 2e-16 ***
stuckanagrafica-b           1.139e-02  6.144e-01   0.019 0.985212
stuckanagrafica-c           1.790e+00  5.388e-01   3.323 0.000891 ***
stuckanagrafica-d           1.721e+00  5.713e-01   3.013 0.002588 **
stuckantiriciclaggio        1.936e+00  5.163e-01   3.749 0.000177 ***
stuckcodice-fiscale         6.693e-01  6.164e-01   1.086 0.277536
stuckconclude               2.464e+01  1.523e+02   0.162 0.871461
stuckcontract-activation    1.885e+00  6.041e-01   3.120 0.001807 **
stuckcontract-subscription  5.182e+00  5.024e-01  10.314  < 2e-16 ***
stuckrapporto               3.910e+00  9.616e-01   4.066 4.77e-05 ***
surv_time                   1.078e-04  6.187e-05   1.743 0.081389 .
extra_attempts              4.785e-02  1.520e-02   3.148 0.001643 **
mean_time                   9.422e-04  3.008e-04   3.132 0.001735 **
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22322.1  on 16101  degrees of freedom
Residual deviance:  8965.6  on 16089  degrees of freedom
AIC: 8991.6

Number of Fisher Scoring iterations: 18
```

Like in case of linear regression, we should check for multicollinearity in the model. As seen in Table 1 below, all X variables in the model have VIF below 4.

Table 1: VIF for the model

	GVIF	Df	GVIF$^($1/(2 * Df))
stuck	1.083490	9	1.004465
surv$_{t}ime$	2.889278	1	1.699788
extra$_{a}ttempts$	1.079711	1	1.039091
mean$_{t}ime$	2.858295	1	1.690649

Then, we can analyze the misclassification error, that is the percentage mismatch of predcited vs actuals, irrespective of 1's or 0's. The lower the misclassification error, the better is your model. Given our optimal cutoff value, we obtain a misclassification error of 0.0286.
Finally, we can plot the Receiver Operating Characteristics Curve, that traces the percentage of true positives accurately predicted by a given logit model as the prediction probability cutoff is lowered from 1 to 0. For a good model, as the cutoff is lowered, it should mark more of actual 1's as positives and lesser of actual 0's as 1's. So for a good model, the curve should rise steeply, indicating that the TPR (Y-Axis) increases faster than the FPR (X-Axis) as the cutoff score decreases. Greater the area under the ROC curve, better the predictive ability of the model. We can see this plot in Figure 6.

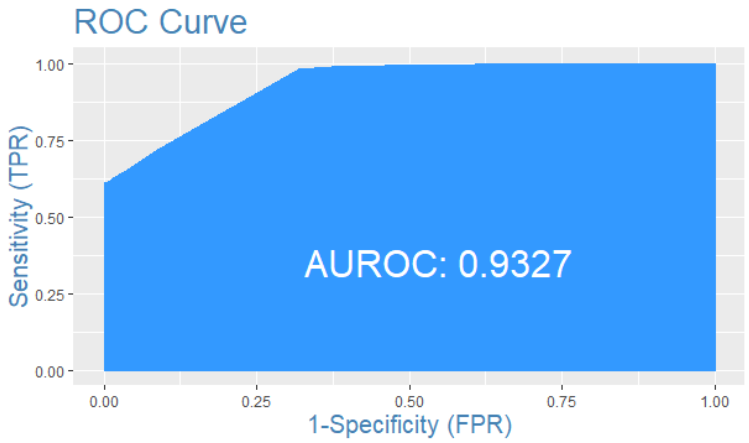

Figure 6: Receiver Operating Characteristics Curve

4 Conclusion

We can analyze the concordance of our estimation which means that the model-calculated-probability-scores of all actual Positive's, (aka Ones) should be greater than the model-calculated-probability-scores of ALL the Negatives (aka Zeroes). Such a model is said to be perfectly concordant and a highly reliable one. In simpler words, of all combinations of 1-0 pairs (actuals), Concordance is the percentage of pairs, whose scores of actual positive's are greater than the scores of actual negative's. For a perfect model, this will be 100%. So, the higher the concordance, the better is the quality of model.
Our model gives the following results:

- Concordance: 0.9308505, i.e 93%;
- Discordance: 0.06914948, i.e. 7%;
- 150912230 number of pairs.

We can also analyze the sensitivity and the specificity of our model. Sensitivity (or True Positive Rate) is the percentage of 1's correctly predicted by the model, while, specificity is the percentage of 0's correctly predicted. We obtain a:

- Sensitivity of 0.6085193, i.e. 60%;
- Specificity of 1, i.e. 100%;

Now, one can ask himself why this model gives us so high precision and why the survival analysis does not give us so strange result. The answer to that is really simple: lack of data. Keep in mind that the omitted data are sensitive ones, so if we really want to deal with this kind of data we should pay attention to the Regulation (EU) 2016/679 (General Data Protection Regulation).
So, basically, we have successfully analyze the survival rate of a customer willing to subscribe to the system along with the creation a model that can help us to know the probability that a certain person will or will not conclude the process, without using any sensitive information about the users.