Survival and Longitudinal Data Analysis

10/11/2022

Project

Frequent employment turnover can create a major loss in the company. We want to predict an employee's risk of quitting the company, for example, within a year. To do this, we will compare survival analysis methods (Cox models, survival random forests, etc.) to classification methods. To compare performance, we will spare 25% of the data as a test sample (be careful to stratify well).

The dataset `tunover2.csv` data contains the following variables:

Name	Type	Description
duration	numeric	experience in months
event	numeric	Censorship flag: 1 if quit, 0 otherwise
gender	factor	gender
age	numeric	age in years
industry	categorical	employee's industry
profession	categorical	employee's profession
traffic	categorical	how employee came to the company
coach	categorical	presence of a coach on probation
head_gender	categorical	gender of the supervisor
greywage	categorical	whether the salary is fully registered with tax authorities
transport	categorical	employee's means of transportation
extraversion	numeric	extraversion score
indepedent	numeric	independent score
selfcontrol	numeric	selfcontrol score
anxiety	numeric	anxiety score
novator	numeric	novator score

The code for the `traffic` variable is given as follows:

- advert (direct contact of one's own initiative)
- recNErab (direct contact on the recommendation of a friend, not an employ of the company),
- referal (direct contact on the recommendation of a friend, an employee of the company),
- youjs (applied on a job site),
- KA (recruiting agency brought),
- rabrecNErab (employer contacted on the recommendation of a person who knows the employee),
- empjs (employer reached on the job site)

1. Import data, check variable types and make necessary changes.

2. Check if data contains `NA` or duplicate lines.

3. Make a histogram for the variable `duration` by coloring according to the value of `event` and calculate the percentage of censorship in the dataset. What do you notice ?

4. Make histograms for continuous covariates and bar charts for discrete ones.

5. Using `corrplot` library, graphically represent the correlations between covariates (be careful to first transform the data that they are entirely numerical, by creating the corresponding dummy variables).

6. Graphically represent the survival functions in the subgroups defined by the categorical variables. What do you notice?

7. Create an 75/25 partition of data in `train` and `test` samples via the `caret` library. Be careful to stratify well on the censorship variable and check that there is approximately the same percentage of censorship in `train` and `test` samples.

8. Make a first Cox model on the `train` sample.

9. Using the `riskRegression` library, represent the Brier score as a function of time. Then code a function that calculates the embedded score on the `test` sample, see https://square.github.io/pysurvival/metrics/brier_score.html for definitions.

10. Repeat the last two questions with a random forest (in the first one, we will take the default parameters) from the library `randomForestSRC`.

11. Which model do you prefer for prediction?

12. Consider an employee whose features are Female of age 30, referred by an employee of the company (referral) in IT industry, profession HR, commuting by bus, having a coach during the probation, with male supervisor, whose characteristic scores are 5 for all categories. Give an estimate of the probability that this employee will stay for longer than 3 years with your best model. Compute estimates for other industry (by changing the profile of the industry only). What do you observe?

13. Now consider another employee with the same profile as above but who has alrady worked for one year. Give an estimate of the probability that this employee will stay for another 2 years with your best model. What is the difference between this probability and the previous probability, in terms of both theory and practice (results)?

14. Try to find an alternative model that improves the performance of the previous ones.